Pattern Recognition & Machine Learning

Lab-8

“Dimensionality Reduction and Feature Selection”

Objectives

Following were the required tasks to be fulfilled in Lab-8 :-

1. Implement PCA (Principal Component Analysis) on a standardized dataset, find out the number of eigenvectors required to preserve 90% of variance, determine the primary contributor amongst elements of 1st eigenvector and plot the transformed dataset using the first 2 eigenvectors..

2. Implement LDA (Linear Discriminant Analysis), compare the details of both, plot the dataset using first 2 linear discriminants and compare the performances of Bayes Classifier from original dataset to transformed dataset using PCA and LDA.

3. Use any of the 2 Feature Selection Methods taught in class on Pre-processed Dataset, identify higher significant features, compute and compare accuracy and f1 scores of any classification model, use Pearson Correlation and show combinations with 70% threshold.

Datasets

Following were the datasets, provided for Lab-8 :-

1. Iris Dataset - data set

2. Diabetes Dataset - dataset

Dependencies

1. Numpy
2. Pandas
3. Sklearn
4. Matplotlib
5. Seabron
6. Mpl_toolkits
7. mlxtend

Preprocessing Methods

Following were some of the techniques/methods explored during Lab-8 :-

1. For Iris Dataset, Labels were converted into numerical classes, using Label Encoder. In addition to this, for standardizing the dataset, Sklearn's StandardScaler function was used, such that Dataset seems to be more distinguishable than before and appear more centralized, with 0 mean (myu) and variance (sigma**2) as 1.

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

2. For Diabetes Dataset, Comparisons were made using Original, Standardized and Min-Maximised Datasets. For the Later one, Sklearn's MinMaxScaler was used as another method to transform the data into simpler form.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Principal Component Analysis

Procedures

1. Standardized the Dataset, such that mean turns out to be 0 and variance as 1.

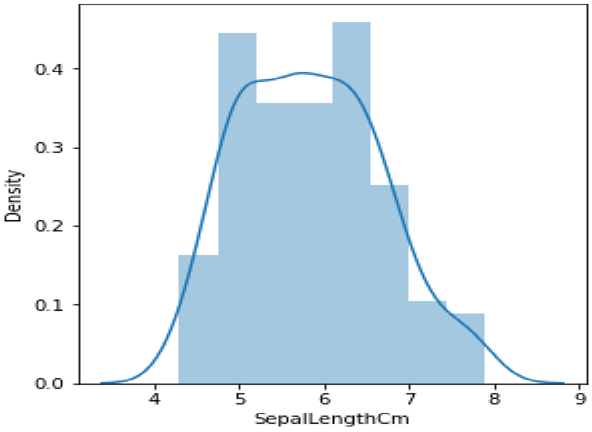
2. Feature-Wise Distributions of Dataset were plotted for Original as well as Standardized Dataset, in order to notice changes more clearly.

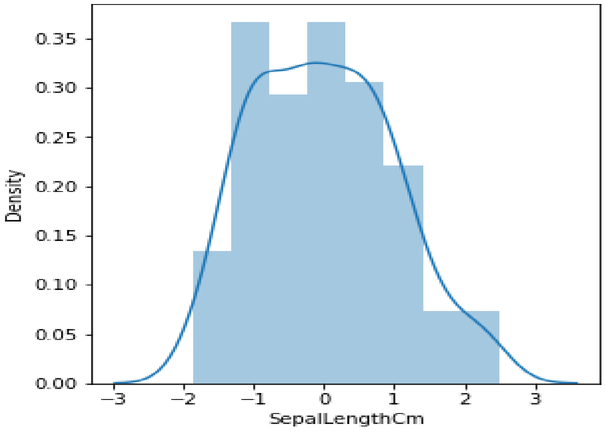
3. Principal Component Analysis was done on Original as well as Standardized Datasets, to observe the effect of Centralizing the Data in the form of results for both the cases.

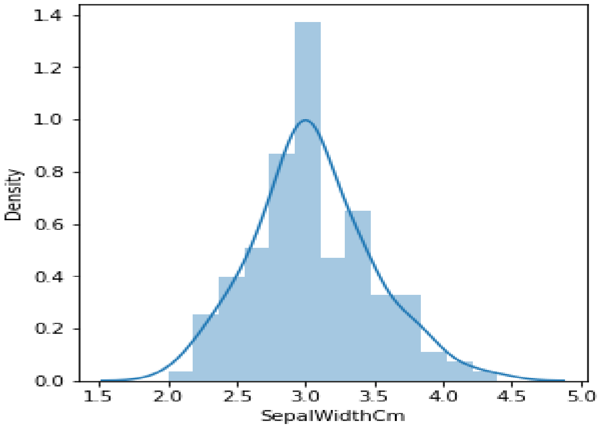
4. Plots were made for both types of Dataset, using the first 3 Principal Components.

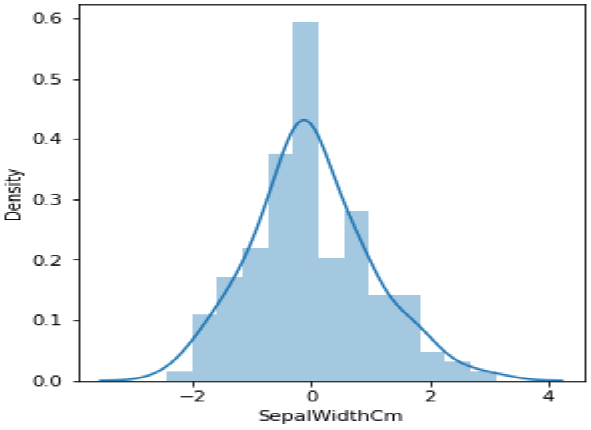
Feature-Wise Distributions

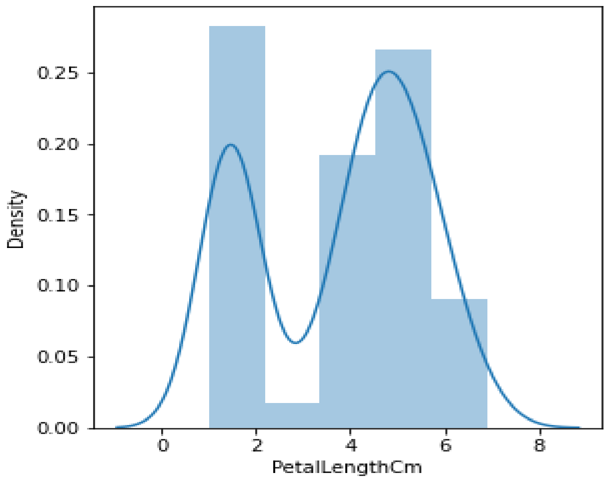
Following were the Plots, before and after standardizing the dataset :-

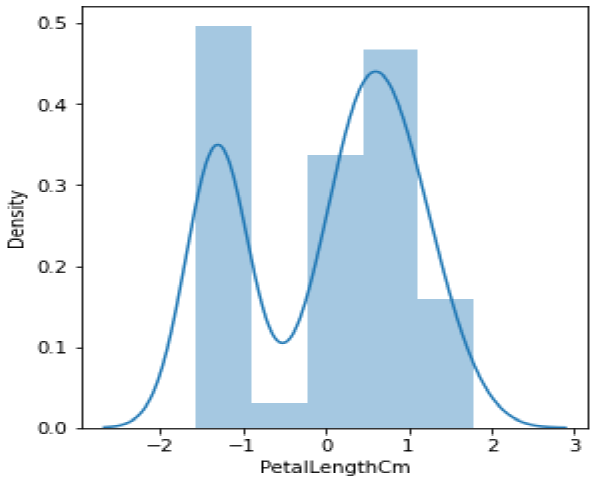


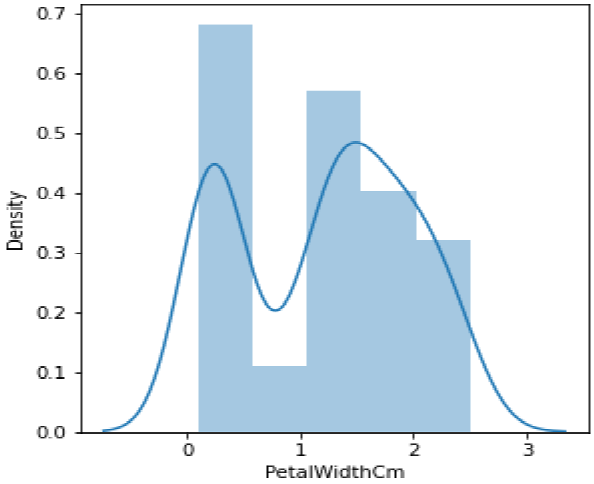


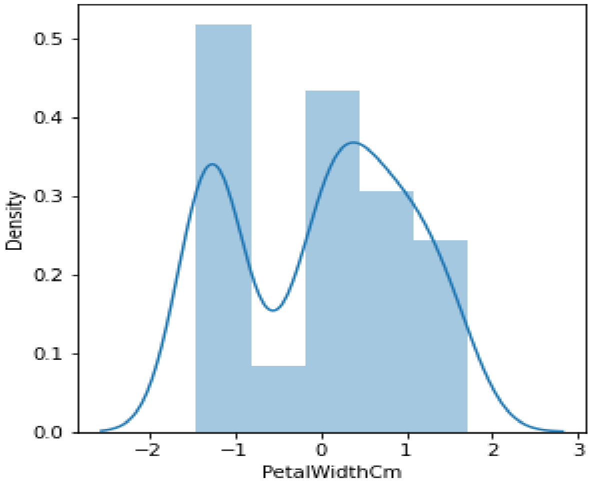


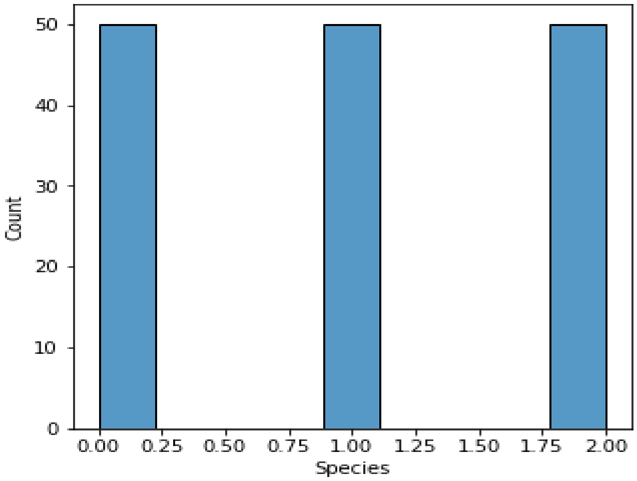



0.25
0.20
0.15
0.10
0.05
0.00
Density
0
2
4
6
8
PetalLengthCm


0.5
0.4
0.3
0.2
0.1
0.0
Density
−2
−1
0
1
2
3
PetalLengthCm

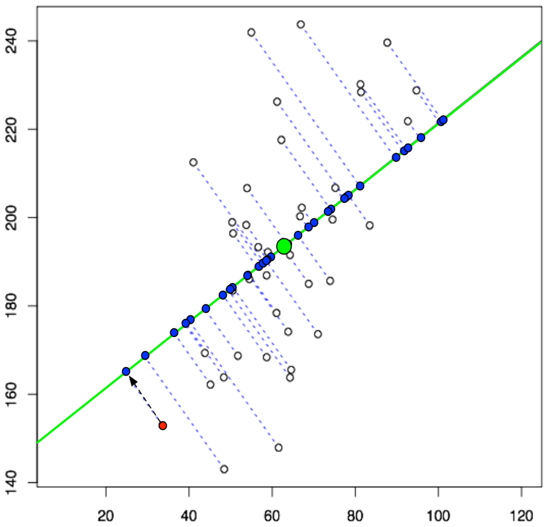

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
0
1
2
3
PetalWidthCm


0.5
0.4
0.3
0.2
0.1
0.0
Density
−2
−1
0
1
2
3
PetalWidthCm


50
40
30
20
10
0
Count
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Species

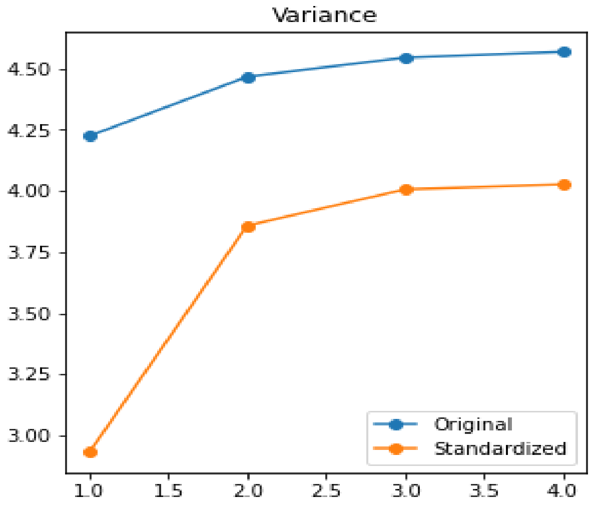
Principal Component Analysis

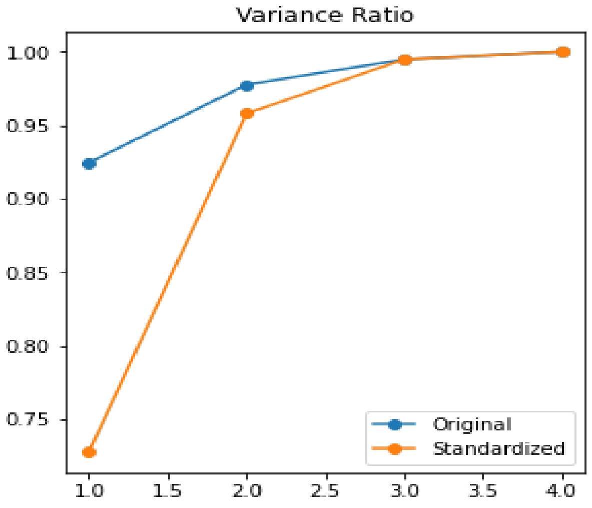
For n-dimensional dataset, Principal Component Analysis converts it into k-dimensional dataset, provided that k is less than n and orthogonal distances between data points & k-dimensional hyperplane should be minimum. But, for more effective results, PCA focuses on maximising the variance (distance between data points and origin), such that transformed data-points seem to be more distinguishable and centralized.



Variance Differences between Standardized and Original Datasets

Following were the plots, obtained while calculating sum of Variances and Variance Ratios :-

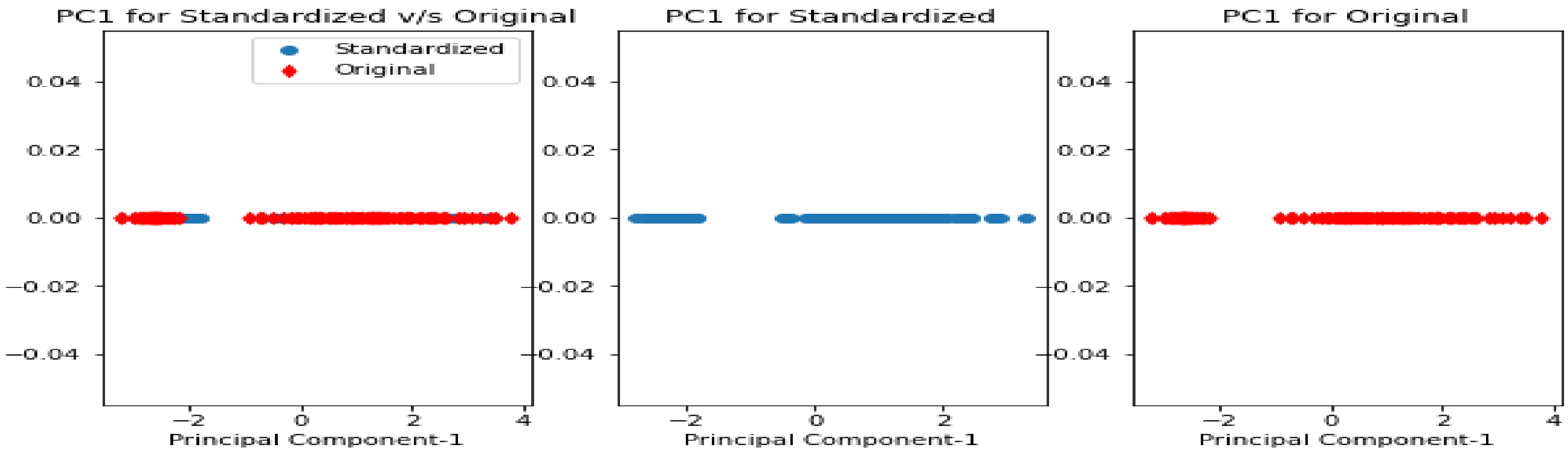




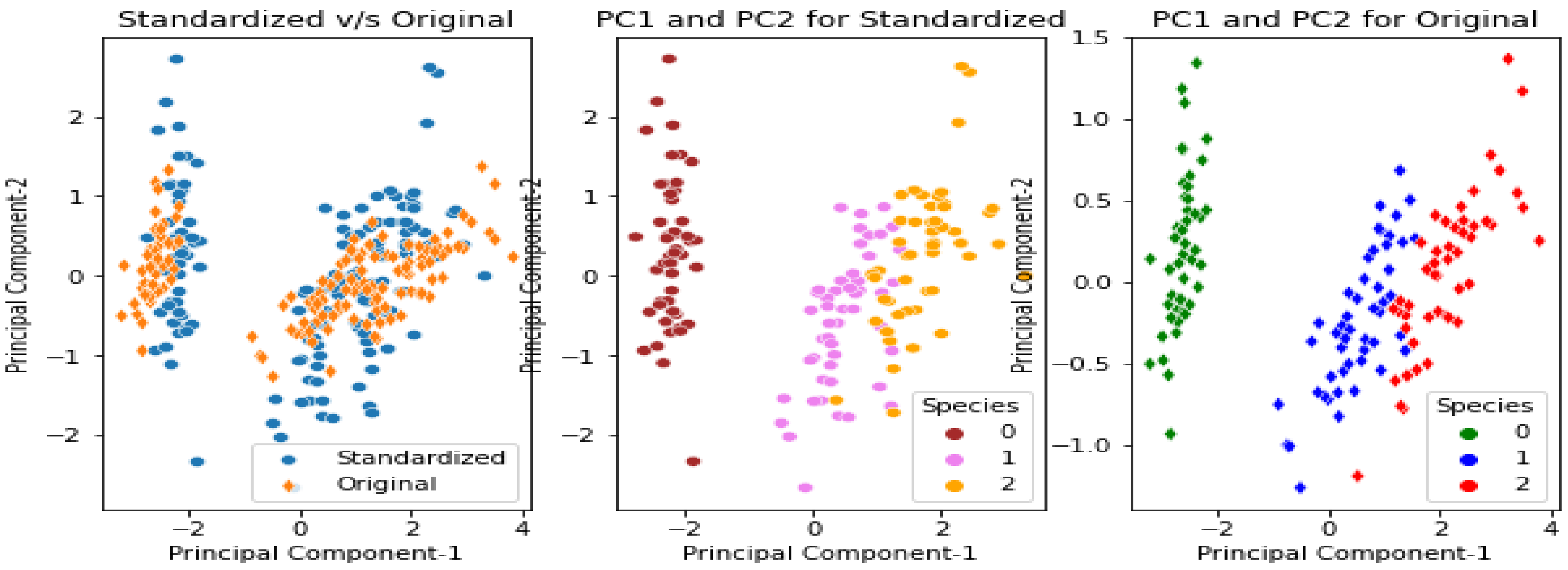
Transformed Dataset Plots

Following were the plots of transformed dataset, from Original as well as Standardized Datasets, using first 3 Principal Components :-

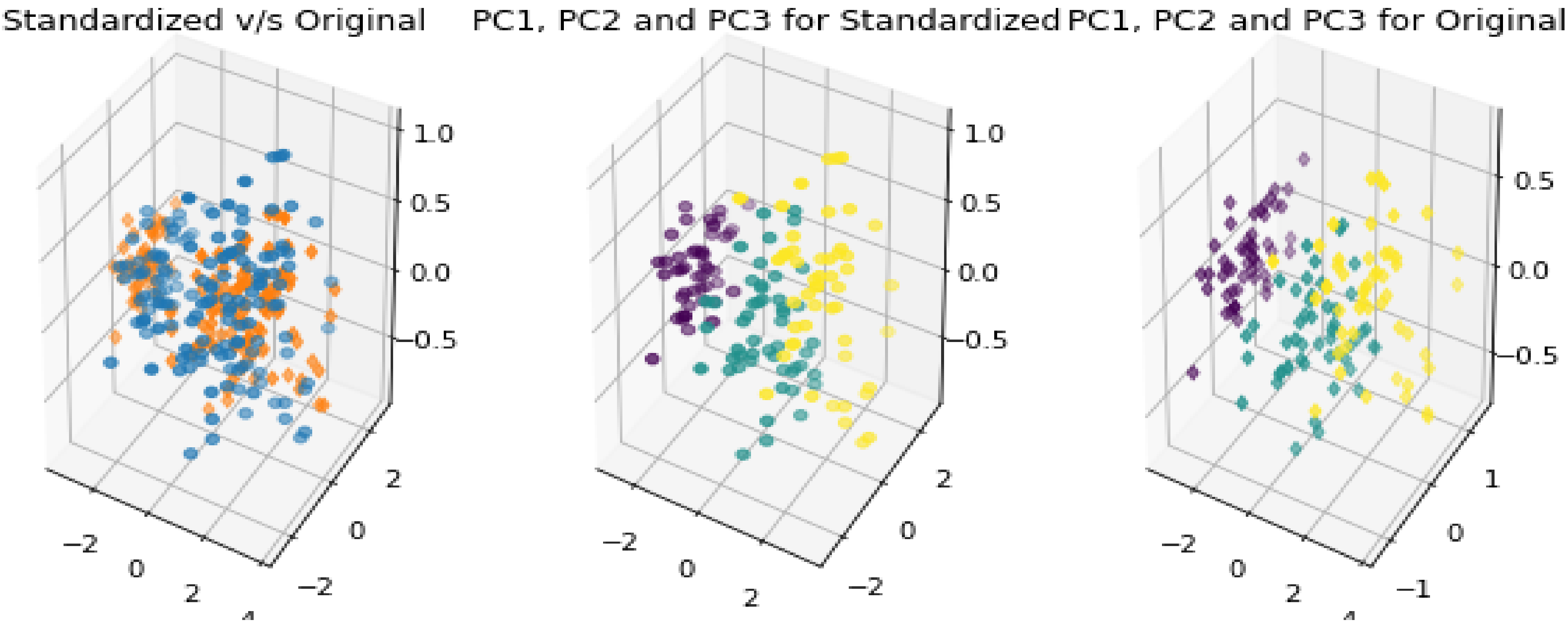
1. Using 1st Principal Component only.



2. Using 1st and 2nd Principal Components



3. Using first 3 Principal Components


Standardized v/s Original

PC1, PC2 and PC3 for Standardized

PC1, PC2 and PC3 for Original

Linear Discriminant Analysis

Procedures

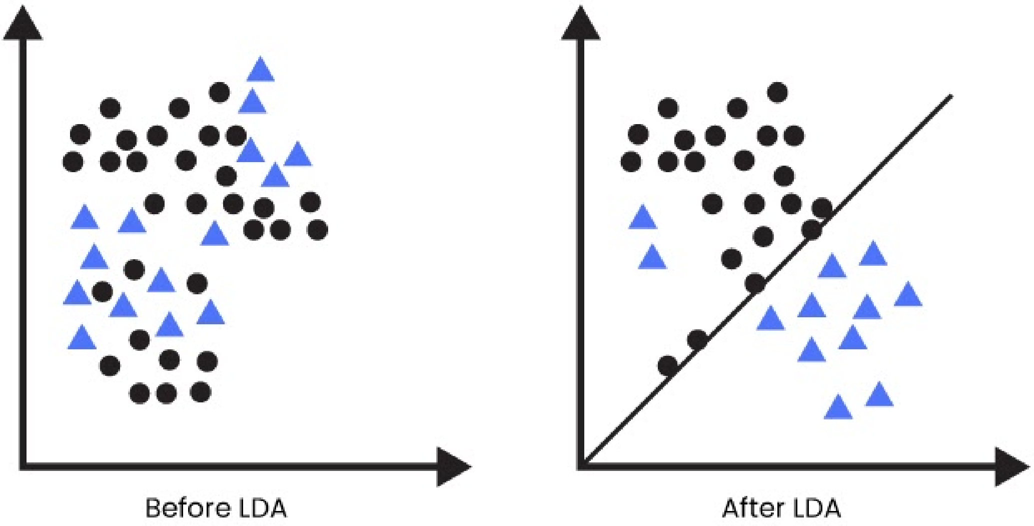
1. Linear Discriminant Analysis was done on Original as well as Standardized Datasets, to observe the effect of Centralizing the Data in the form of results for both the cases.

2. Plots were made for both types of Dataset, using the first 2 Linear Discriminants and compared with the results of Principal Components.

3. A Bayesian Classifier was trained for Dataset and their transformed versions. Performances were compared for the model, implemented on each and every form of dataset.

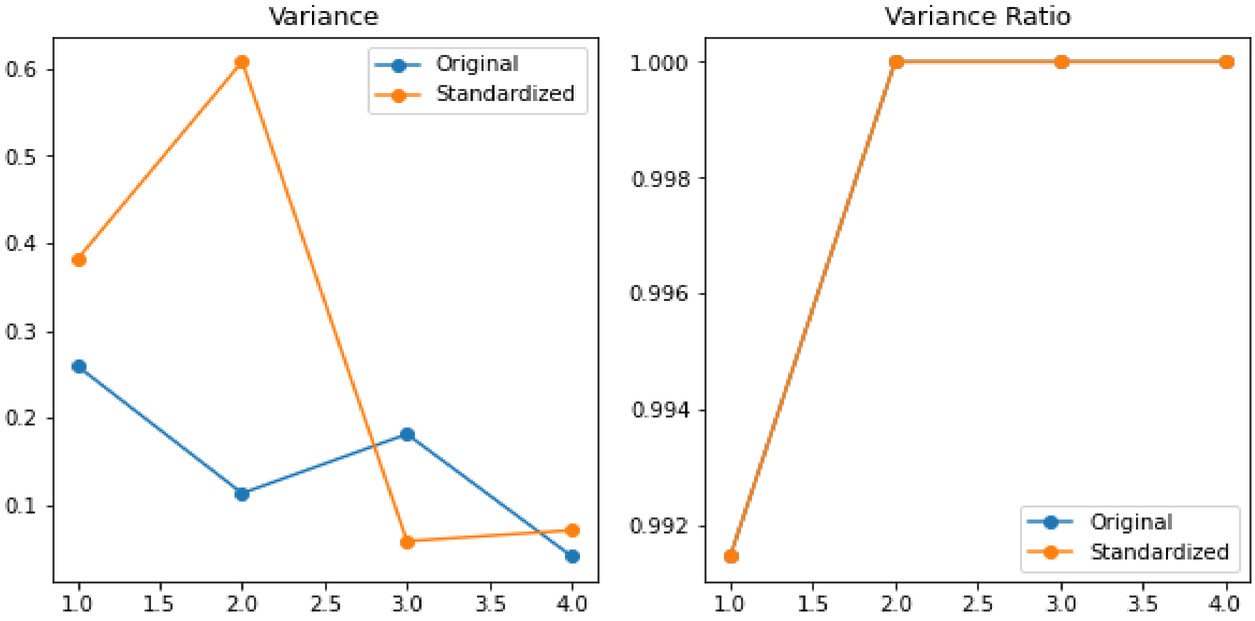
Linear Discriminant Analysis

For n-dimensional dataset, Linear Discriminant Analysis converts it into k-dimensional dataset, provided that k is less than n and intra-cluster distances should be minimum. But, for more effective results, LDA focuses on maximising the inter-cluster distance such that transformed data-points seem to be more distinguishable. Infact, Linear Discriminant Analysis uses Supervised Learning Approach, to transform the data points.



Variance Differences between Standardized and Original Datasets

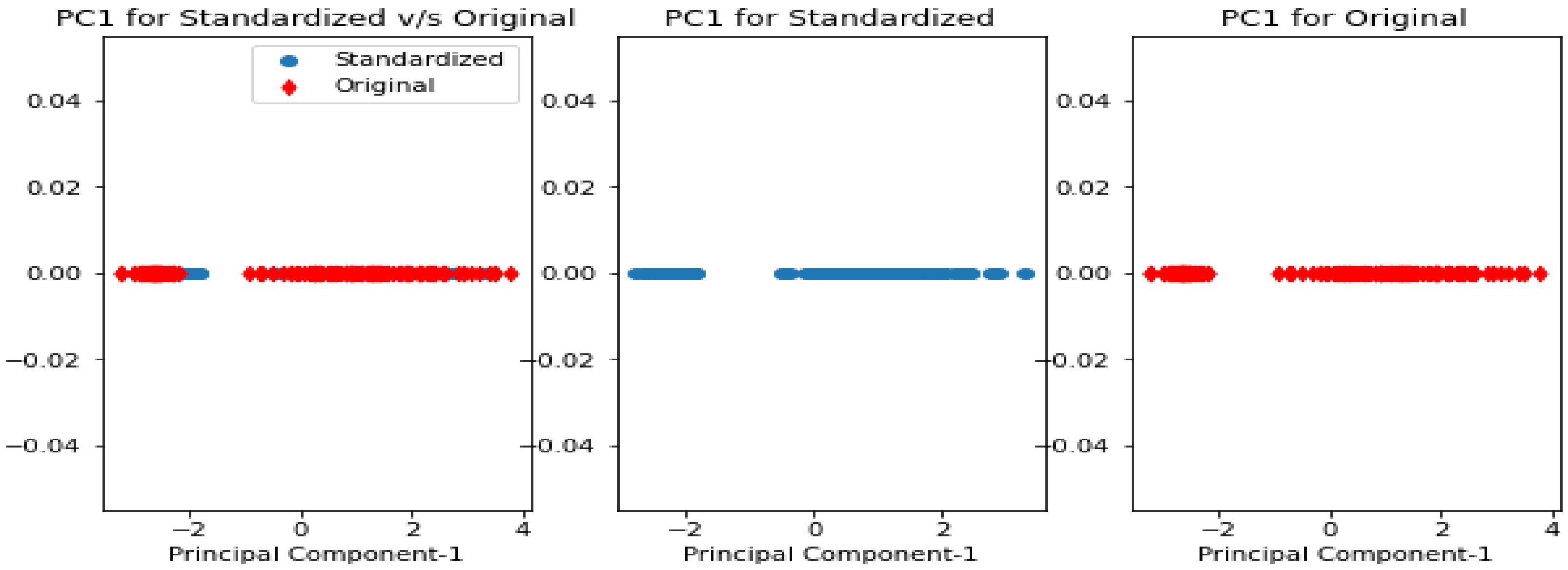
Following were the plots, obtained while calculating sum of Variances and Variance Ratios :-

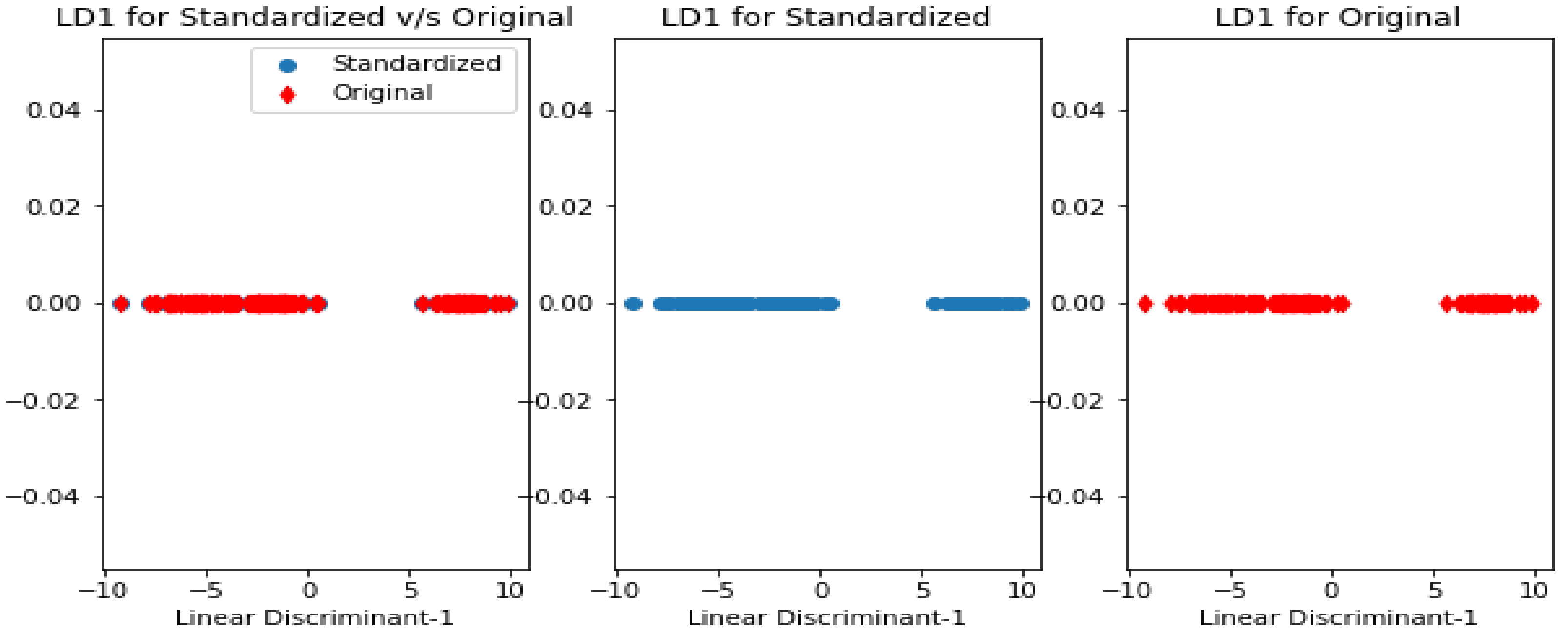


Comparison between PCA and LDA plots

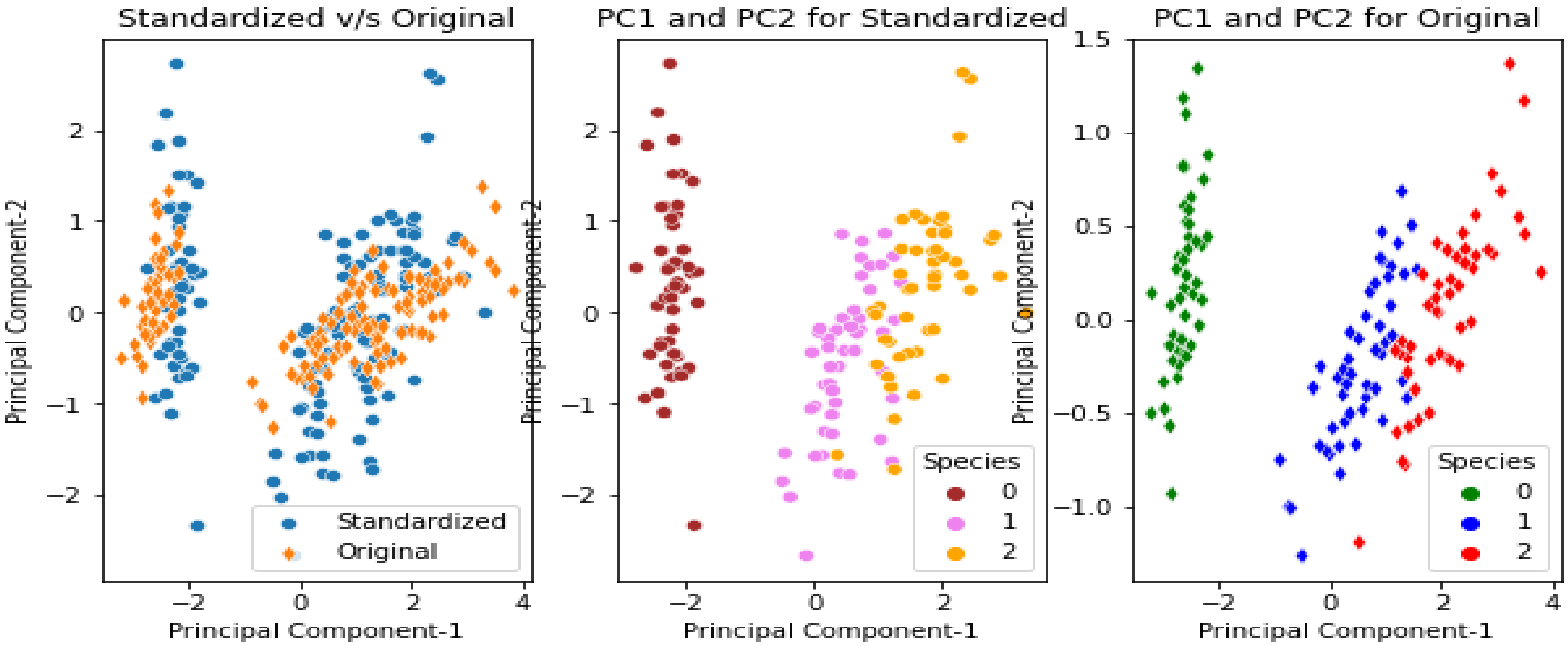
Here are the comparative plots between Principal Component Analysis and Linear Discriminant Analysis, as follows :-

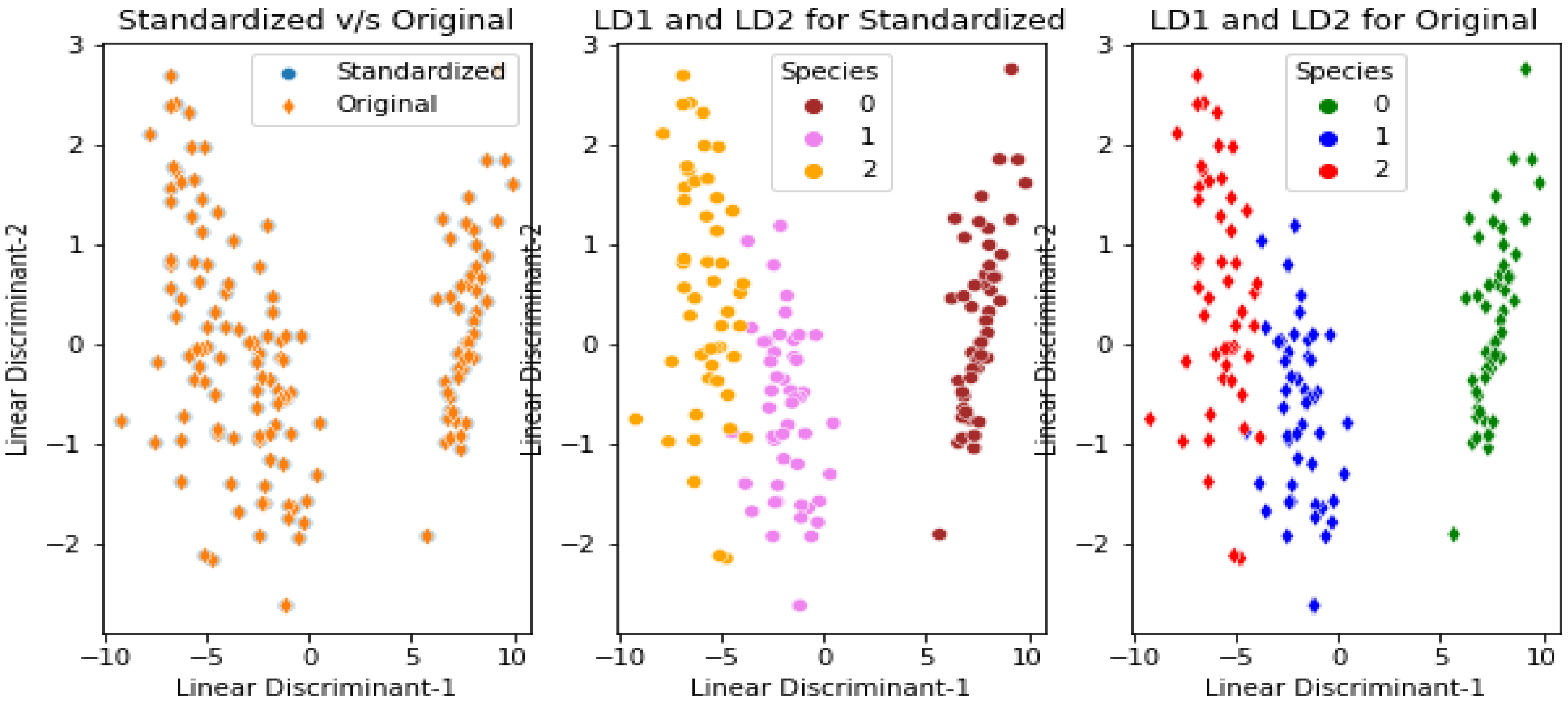
1. Using Single Principal Component and Linear Discriminant





2. Using Couple of Principal Components and Linear Discriminants





Effects of PCA & LDA on Bayesian Classification Task

Here were the Following details, regarding the implementation of Bayes Classifier on Iris Dataset :-

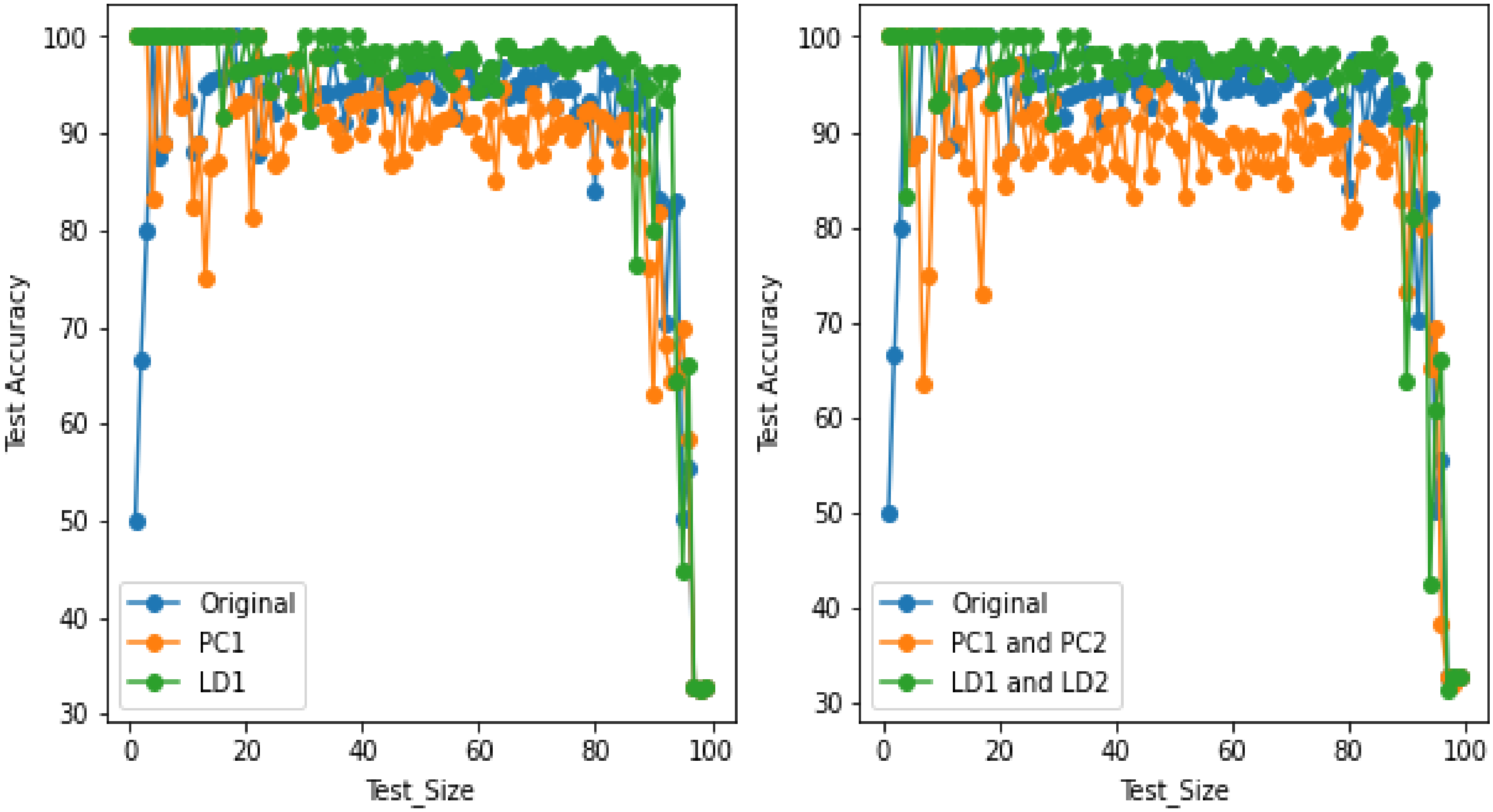
1. Gaussian Naive Bayes Model was chosen, amongst other variants.

2. For 100 different train-test-splits of each type of Dataset, Model was trained and then run for predicting the test portions. Their Accuracies were plotted and compared with reference plot.

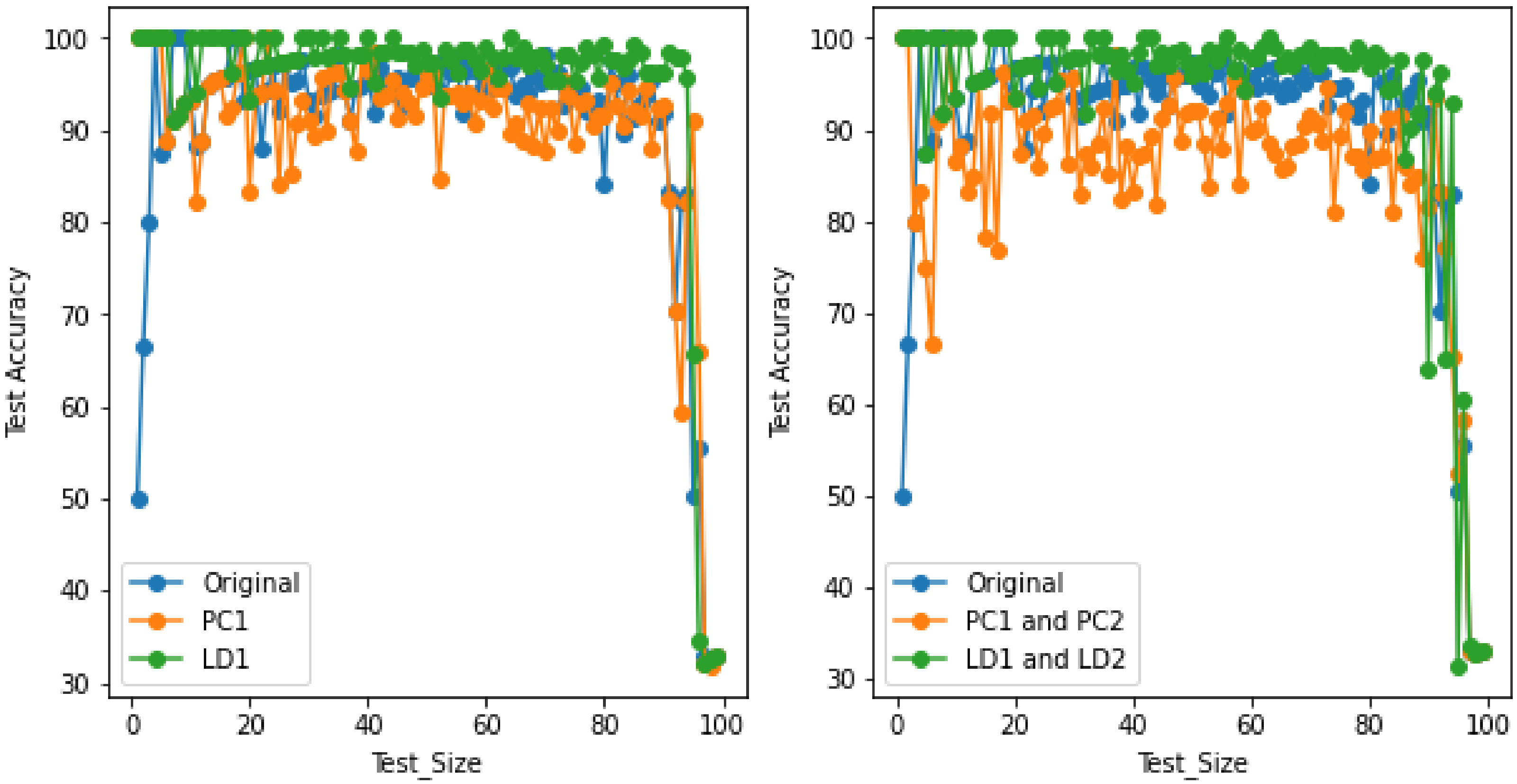
3. Plots were made uptil 2 Principal Components and 2 Linear Discriminants.

Following were the Test_size v/s Accuracy Plots for observing the effects of PCA and LDA on Bayesian Classification Task :-

1. Original Dataset v/s Standardized Principal Component & Linear Discriminant



2. Original Dataset v/s Original Principal Component & Linear Discriminant



Feature Selection Techniques

Procedures

1. Feature-Wise Distributions and their descriptions , regarding 3 types of datasets,i.e Original, Standardized and Min-Maximised, were compared to each other.

2. Any 2 types of Feature Selection Techniques were used to determine the main 5 features, out of 8 features, for easier computation and thereby trying to improve the accuracy of the model.

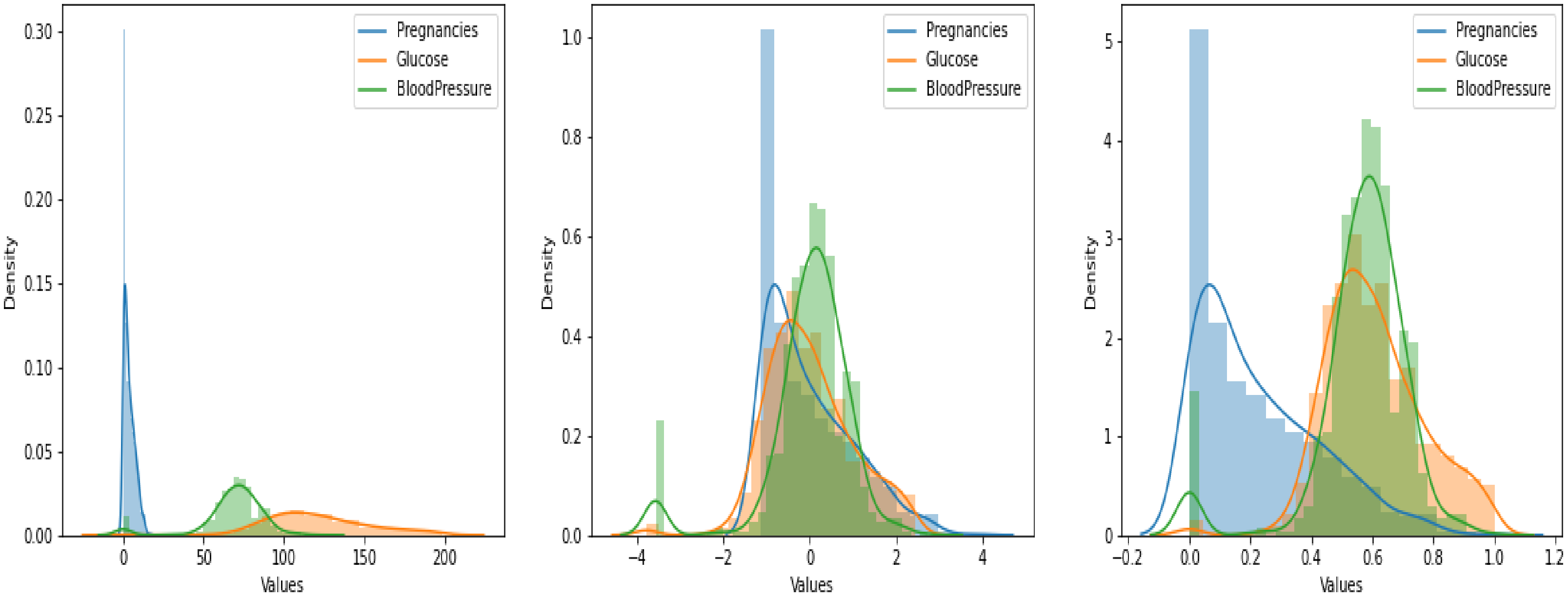
3. Accuracy and F1 Scores were compared for each type of feature selection technique and each type of dataset.

4. Using Pearson Correlation and 70% thresholding, important feature mappings were brought out.

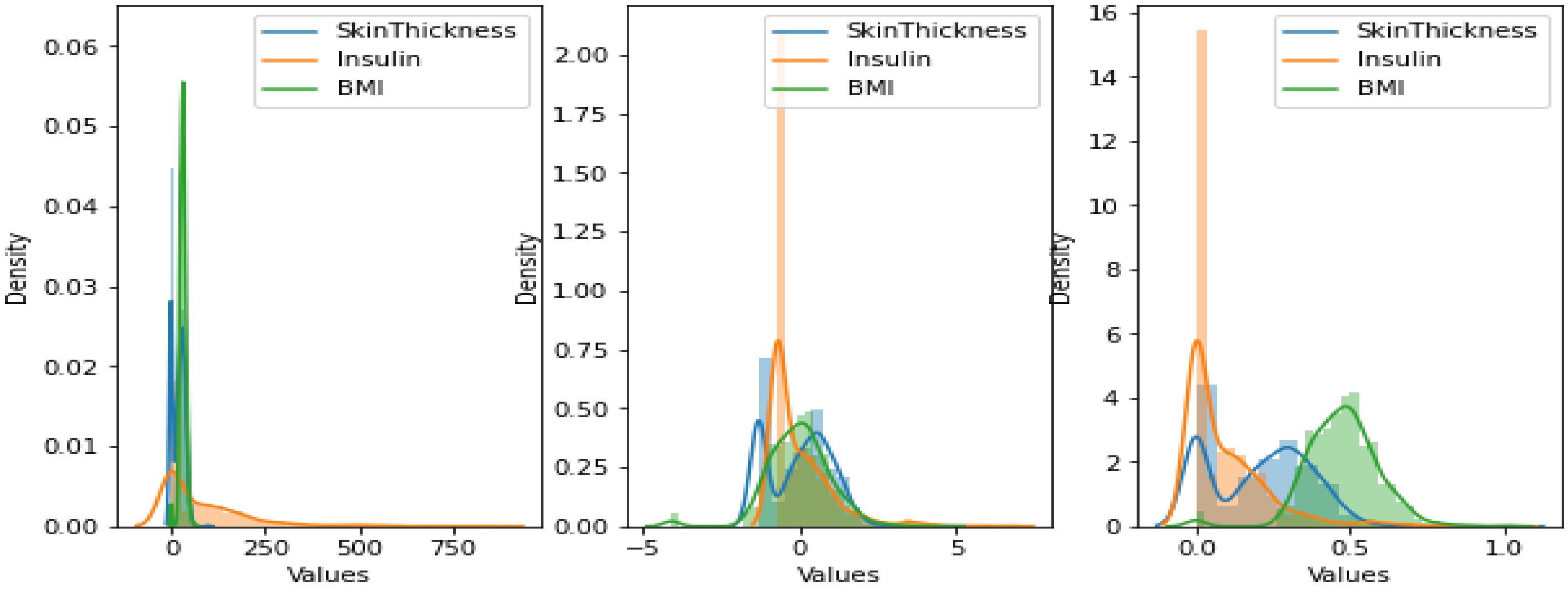
Feature-Wise Distribution Plots

Following were the Feature-Wise Distributions for Original, Standardized and Min-Maximised Datasets :-

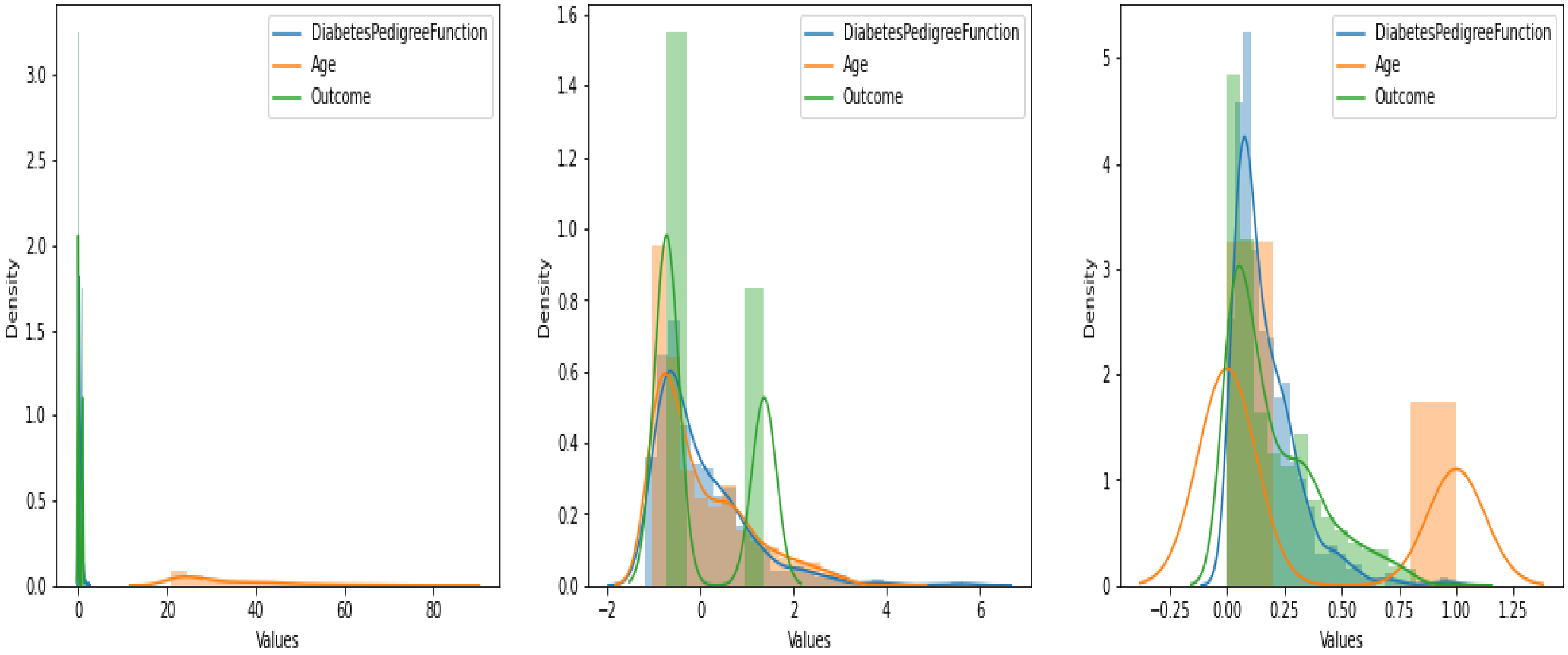
1. Pregnancies, Glucose and BloodPressure



2. SkinThickness, Insulin and BMI



3. DiabetesPedigreeFunction, Age and Outcome



Feature Selection Methods

Following were the 2 methods, incorporated for selecting main features amongst 8 total features of Diabetes Dataset :-

1. Forward Feature Selection :-

   This type of Feature Selection chooses the best feature, amongst other features, according to some arbitrary criterion function, adds to the desired subset, till the number of elements in the subset becomes equal to that mentioned by the User (Priori).

2. Exhaustive Feature Selection :-

   This type of Feature Selection uses the Brute Force approach, constructs sets of combinations, according to minimum and maximum number of features specified by a user, then it chooses the best combination set, by trial running each.

Insights about Classification Model and Features

Regarding the configuration of the Classification Task, “Support Vector Machine with Polynomial Kernel of degree 2” was used, with a train-test ratio of 7:3. Following was the table obtained after applying feature selection techniques :-

Type of Dataset	Forward Feature Selection List	Exhaustive Feature Selection List
Original	['Pregnancies', 'Glucose', 'BloodPressure', 'Insulin', 'BMI']	['Pregnancies', 'Glucose', 'BloodPressure', 'Insulin', 'BMI']
Standardized	['Pregnancies', 'Glucose', 'Insulin', 'BMI', 'Age']	['Pregnancies', 'Glucose', 'SkinThickness', 'Insulin', 'BMI']
Min-Maximised	['Glucose', 'BloodPressure', 'BMI', 'DiabetesPedigreeFunction', 'Age']	['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'DiabetesPedigreeFunction']

Test - Accuracies

Following were the Accuracies obtained after applying feature selection techniques, compared to original features :-

Type of Dataset	Accuracy, considering All Features	Accuracy, considering Features from Forward Selection	Accuracy, considering Features from Exhaustive Selection
Original	70.9956%	70.5627%	70.5627%
Standardized	67.5324%	67.0995%	70.5627%
Min-Maximised	77.4891%	79.6536%	78.3549%

F1-Scores

Following were the F1-Scores obtained after applying feature selection techniques, compared to original features :-

Type of Dataset	F1-Score, considering All Features	F1-Score, considering Features from Forward Selection	F1-Score, considering Features from Exhaustive Selection
Original	0.5179	0.5072	0.5072
Standardized	0.3902	0.5209	0.5239
Min-Maximised	0.5937	0.4802	0.4802

Conclusion

Following were the conclusions, drawn from Lab-8 :-

1. There is no effect of Standardization on the results of Linear Discriminant Analysis, whereas the data points differed in Principal Component Analysis. This might be just because PCA uses Unsupervised approach, whereas LDA uses Supervised Approach.

2. For preserving at least 90% of data variance, at least n_components=2, would be required to obtain.

3. For Standardized as well as Original Dataset, the respective 1st eigenvectors were as following:-

   ```
   [ 0.52237162, -0.26335492,  0.58125401,  0.56561105]
   [ 0.36158968, -0.08226889,  0.85657211,  0.35884393]
   ```

   We can say that 3rd Feature (PetalLengthCm) has larger coefficient from others. PetalLengthCm was positively correlated to SepalLengthCm and PetalWidthCm, whereas it was negatively correlated to SepalWidthCm.

4. For Bayes Classification Task, Linear Discriminant Analysis turned out to be performing better than Original Dataset and Principal Components. This might be just because LDA focuses more on building a decision boundary, which pretty much separates the data points according to the classes. This thing, whereas not so prominently seen in PCA, which focused on general distinguishability, through maximizing variance.

5. The Correlation Matrix, formed for Original, Standardized and Min-Maximised Datasets, turned out to be the same, and having exactly the same values for each pair. Pearson Correlation coefficients above 0.7 (70%) were found to be on feature mapping like this :-

   Pregnancies ←----------------------------->Pregnancies
   BMI ←----------------------------->BMI
   Glucose ←----------------------------->Glucose
   Insulin ←----------------------------->Insulin
   BloodPressure ←----------------------------->BloodPressure
   DiabetesPedigreeFunction ←----------------------------->DiabetesPedigreeFunction
   Age ←-----------------------------> Age

6. Depending upon the type of Feature Selection Technique and Type of Preprocessing Techniques, the Main Features can differ from each other and sometimes dominate more than others. Like in Forward Feature Selection, a criterion function decides the

entry of main features, whereas Exhaustive Feature Selection focuses more on Combination Searching.

7. It is totally dependent on the type of pre-processing technique, whether accuracy and f1-score of main features or with all features have higher values.