Pattern Recognition and Machine Learning

Major Project

Report

Team Members	Roll Number	Contribution
Kalbhavi Vadhi Raj	B21EE030	Report , MLP, and SVM
Aneesh Atul Borkar	B21EE079	Data preprocessing, exploratory data analysis, Gaussian NB , Decision tree
Harsh Nawal	B21AI015	Remaining Models

Project 9: Detecting Parkinson's Disease

Abstract:

This study aimed to construct a supervised learning model for classifying medical subjects into two groups based on their Parkinson's disease status. The dataset comprises a variety of audio parameters extracted from voice recordings of patients. The dataset is skewed, as 23 of the total 31 patients in the recording are positive. As a result, we used both accuracy and the F1 score as measures. We've employed dimensionality reduction and feature selection techniques and then trained multiple models on them.

All the steps taken are:

1) Data Pre-processing
2) Exploratory data analysis
3) Model implementation
4) Evaluation of the model
5) Hyperparameter tuning
6) Final Evaluation of the Model

Introduction:

In this investigation, we attempted to categorise patients as either healthy or sick using a variety of supervised learning algorithms. Initially, we employed linear discriminant analysis (LDA) to determine whether or not the data were linearly separable. Then, we utilised principal component analysis (PCA) with naive Bayes classification to determine the efficacy of this method.

Then, we attempted the sequential forward feature selection algorithm with the Naive Bayes classifier as the foundational model.

Then, we attempted to identify the optimal feature using the sequential forward feature selection algorithm and the Decision Tree classifier as the base model. On the resulting datasets, we then evaluated the precision of various models.

The various models used in this project are:

1. Gaussian NB
2. Decision Tree Classifier
3. Bagging with the Decision Tree Classifier as the base ensemble
4. AdaBoost with the Decision Tree Classifier as the base ensemble
5. Xgboost Classifier
6. Neural Network
7. Support Vector Machine
8. KNN Classifier

All the above-mentioned models are trained on four different datasets simultaneously: standard data, PCA data, LDA data, and SFS data.

Then we compared the performance of the models on each of the datasets.

Brief Description of the Features:

Feature Name	**Description**
name	ASCII subject name and recording number
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Exports of goods and services per capita. Given as %age of the GDP per capita
MDVP:Flo(Hz	Total health spending per capita. Given as %age of GDP per capita
MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Several measures of variation in fundamental frequency
MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA	Several measures of variation in amplitude
NHR, HNR	Two measures of the ratio of noise to tonal components in the voice
RPDE, D2	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1,spread2,PPE	Three nonlinear measures of fundamental frequency variation
status	The health status of the subject (one) - Parkinson's, (zero) - healthy

1) Data Pre-Processing

After removing nil/na values, we standardised the data. Then we did outlier analysis on the data and manually removed outliers.

Outlier Analysis:

We found the correlation between all the features in the data and selected the feature pairs with the highest correlation, as these features can make it easier to visualise the outliers.

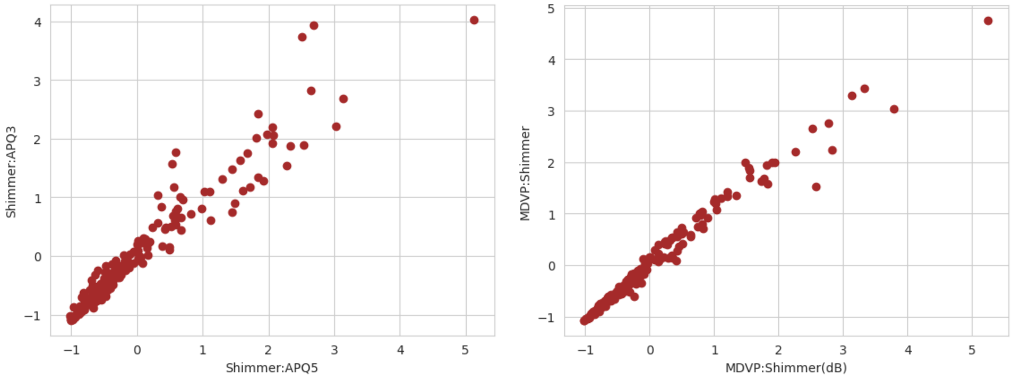


Fig. 1: Some of the plots generated during outlier analysis.

From the plot, we can clearly see that there is an outlier when MDVP:Shimmer (dB) is greater than 4. Also, when shimmer:APQ5 is greater than 4.

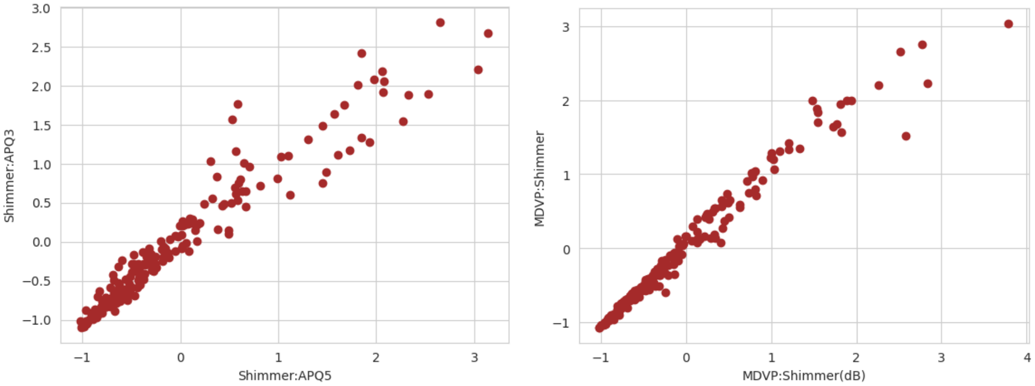


Fig. 2: Same plots after removing outliers.

Dimensionality Reduction:

Applying PCA to the Data:

Initially, we applied PCA to the data without adding a value for k in order to estimate the proportion of variance contributed by each feature.

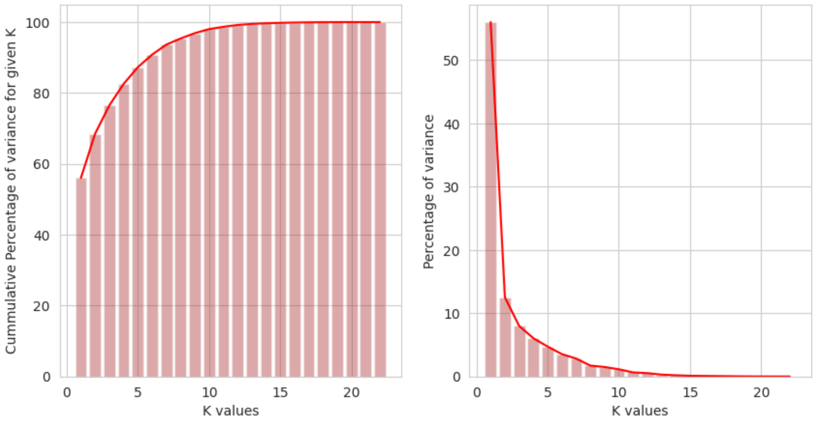


Fig. 3: Plots explaining the covariance conserved by each component of PCA

So from the above graph, we decided to take the best 15 PCA features, as they conserved nearly 100 percent (99.7912%) accuracy.

Applying LDA to the Data:

After applying LDA to the data, we get only one feature, as the number of classes in this dataset is 2.

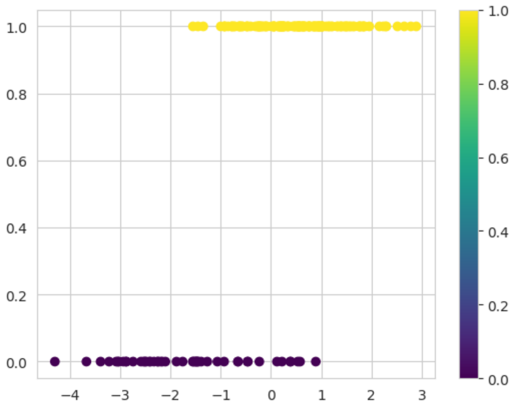


Fig. 4: Plot of LDA data corresponding to labels and colour coded with respect to class

We can see that we can fit the sigmoid function into this plot. However, the relationship will not be linear.

Then we split the data into three parts in the ratio 7:2:1 for train, val, and test.

2) Exploratory Data Analysis

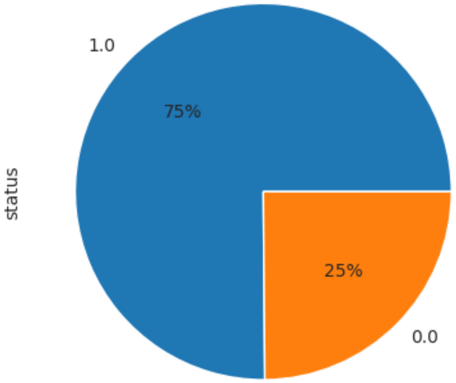


Fig. 5: Pie chart representing the status of the patient

From the above pie chart, we can conclude that only ¼ of the patients are healthy, whereas others are ill.

Plotting histograms and density curves:

For original data:

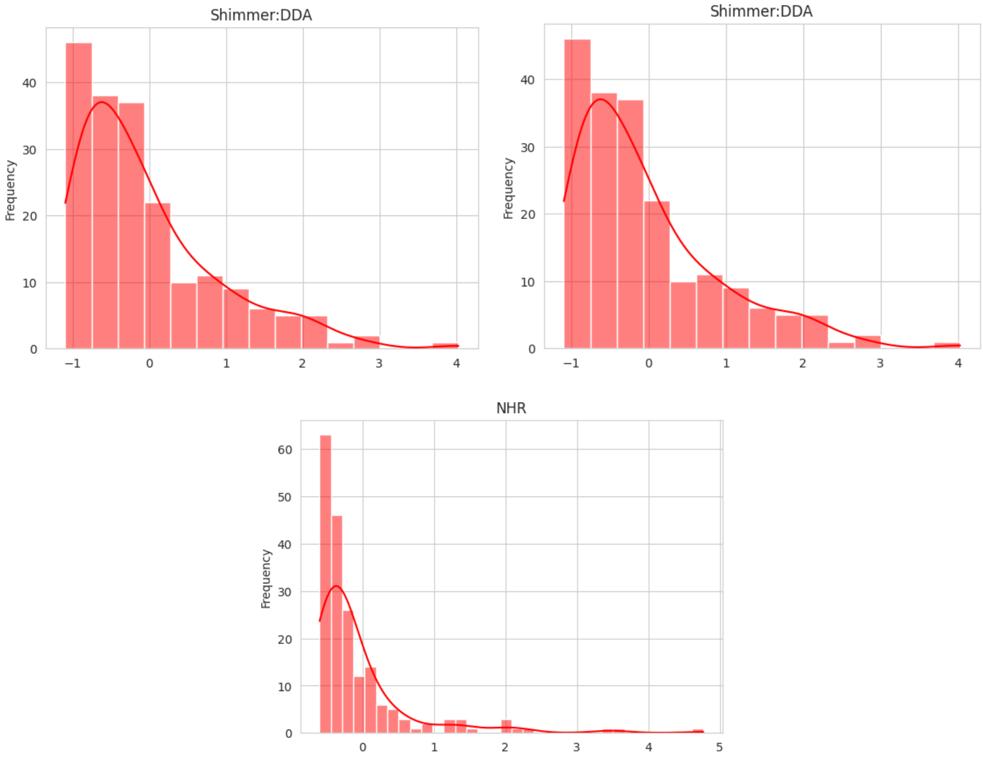


Fig. 6: Histogram of various features in the standardised data along with their estimated density curve

We can see from the plots that none of these features have a gaussian distribution.

For PCA data:

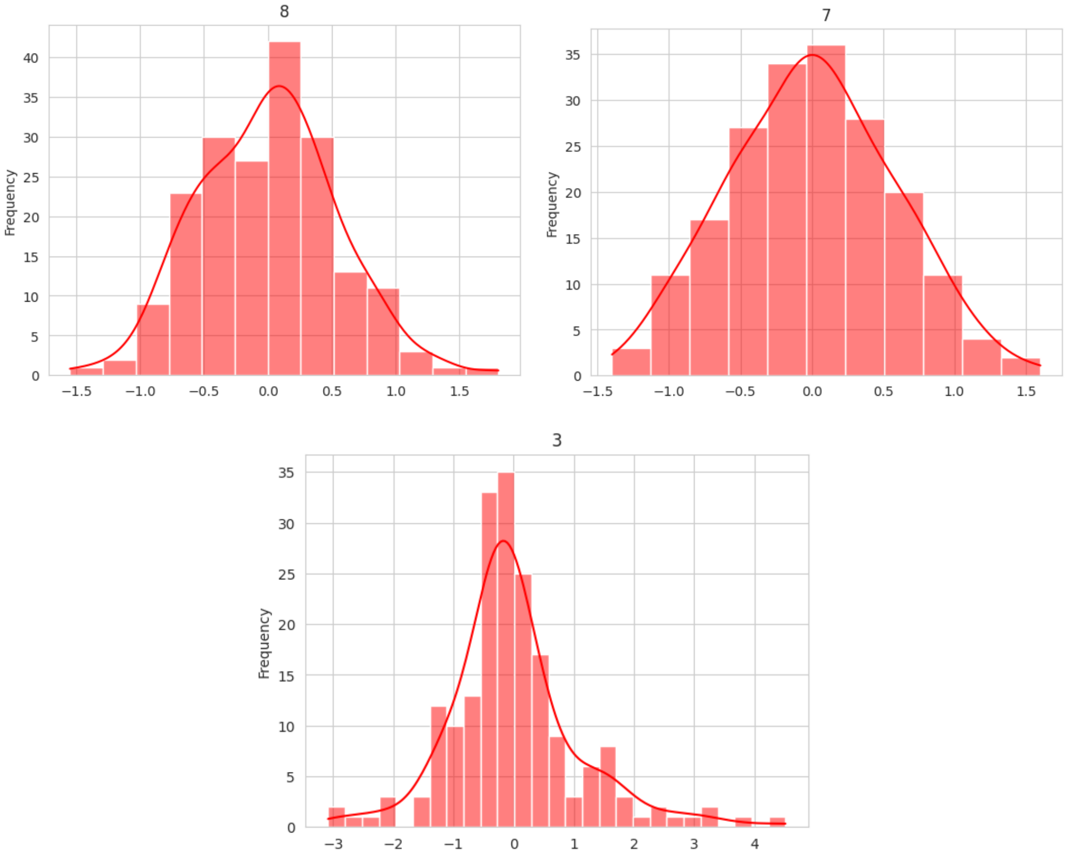


Fig. 7: histogram of various features in the PCA data along with their estimated density curve

From the above graphs, we can conclude that the features we obtained through PCA are very similar to gaussian distributions.

For LDA data:

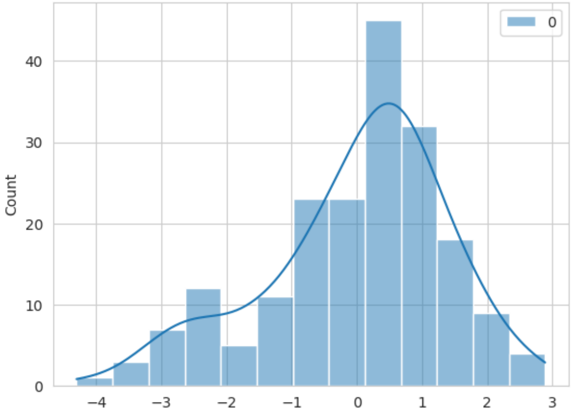


Fig. 8: histogram of a feature in LDA data along with its estimated density curve

We can see that for LDA, its feature distribution is a bit skewed towards the +ve x-axis.

Heatmaps:

For standardised data:

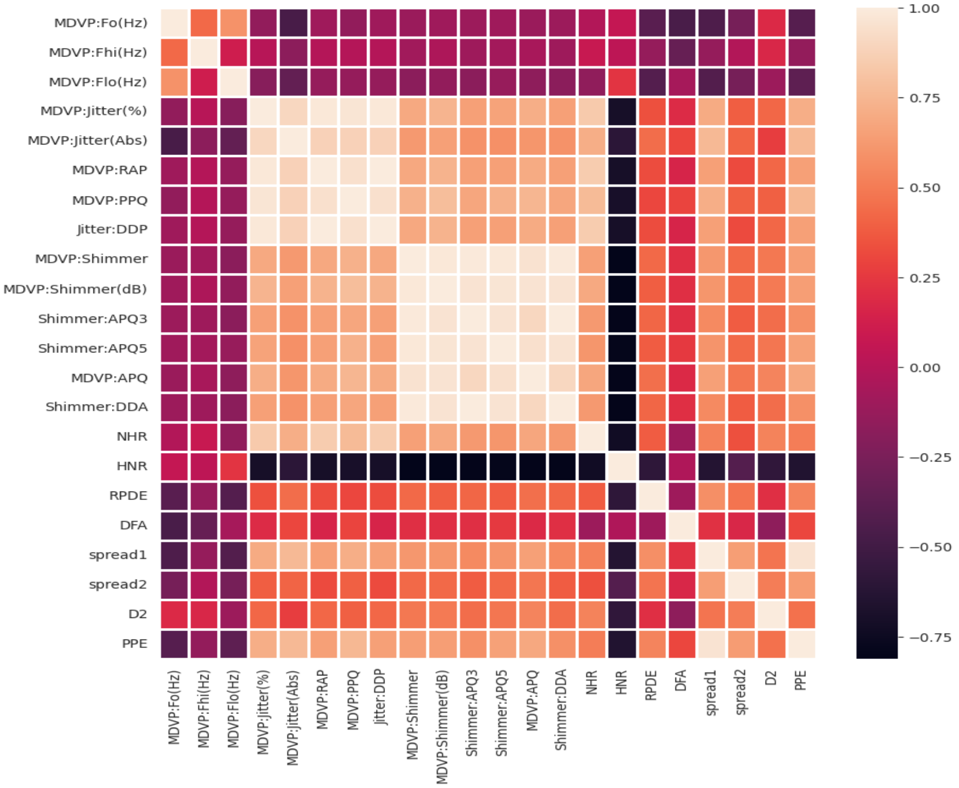


Fig. 9: heatmap of the correlation between the features of the original data

For PCA data:

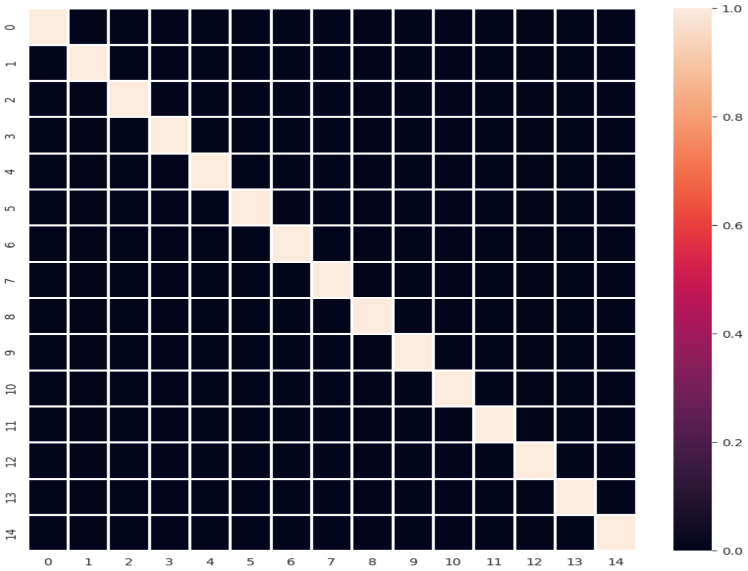


Fig. 10: heatmap of the correlation between the features of the PCA data

We can see that all features are independent of each other, as their correlation is zero.

3) Model Implementation

Gaussian Naive Bayes:

First, we selected the best features using the sequential feature selection algorithm by using Gaussian NB as the base estimator.

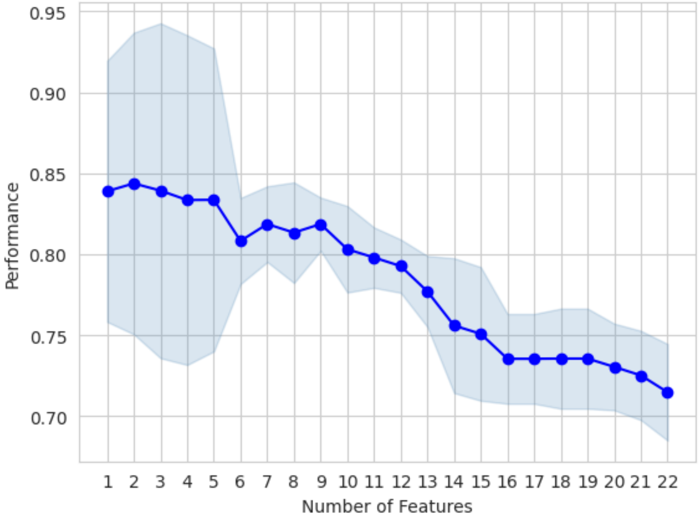


Fig. 11: Performance vs. number of features for the SFS algorithm

It is evident from the above plot that the best value of number features is two, and these two features are selected to form SFS data.

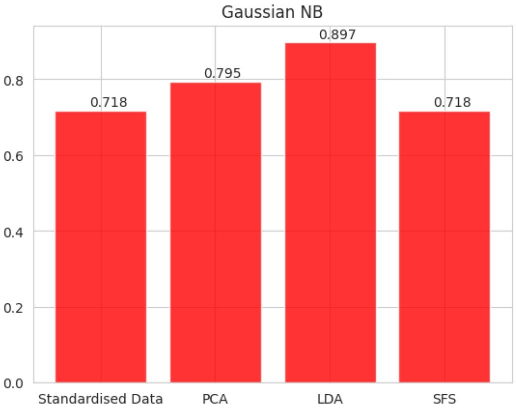


Fig. 12: Plot of accuracy vs. dataset used for training and testing

If we exclude the LDA data, we can see that we get the highest accuracy with PCA, which makes sense as in the exploratory data analysis we saw that all the features of PCA data are independent and resemble a gaussian distribution.

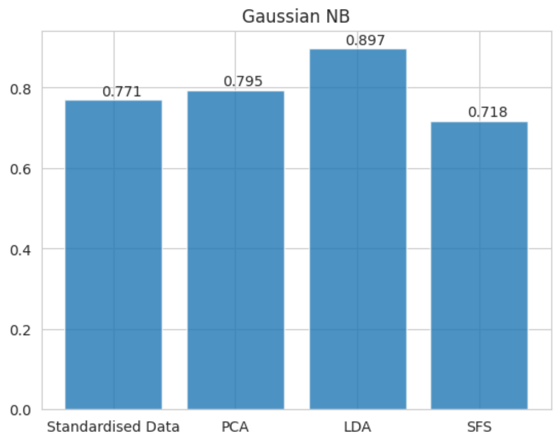


Fig. 13: Plot of F1 score vs. dataset used for training and testing

From the F1 score, we can also draw similar reasoning for the high value of the F1 score on PCA data.

Decision Tree Classifier:

Firstly, we selected the best features using the sequential feature selection algorithm by using the Decision Tree as the base estimator.

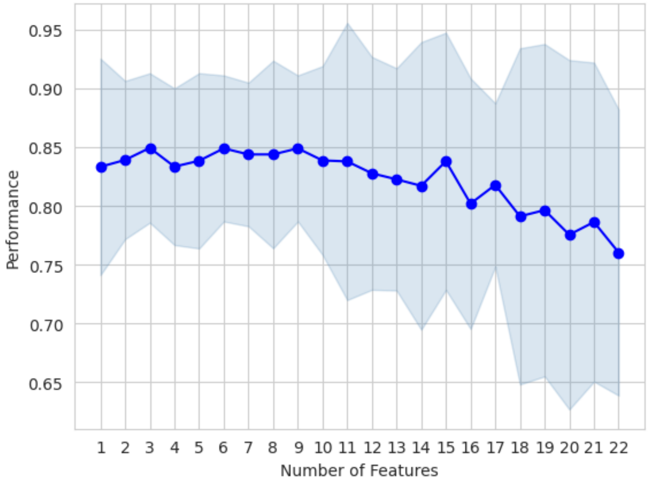


Fig. 14: Performance vs. number of features for the SFS algorithm

It is evident from the above plot that the best value of number features is three, and these three features are selected to form SFS data.

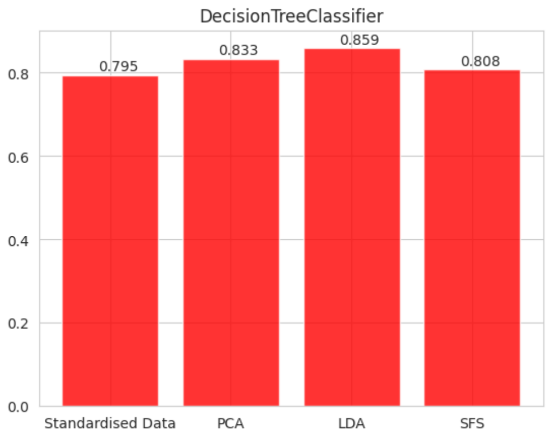


Fig. 15: Plot of accuracy vs. dataset used for training and testing

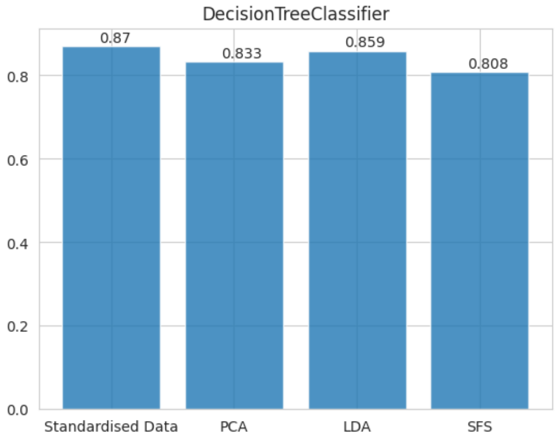


Fig. 16: Plot of F1 score vs. dataset used for training and testing

Hyper-Parameter Tuning:

To increase the accuracy, we tried varying the values of max_depth and chose the value that gave the highest accuracy.

Performance with tuned Hyper- Parameters:-

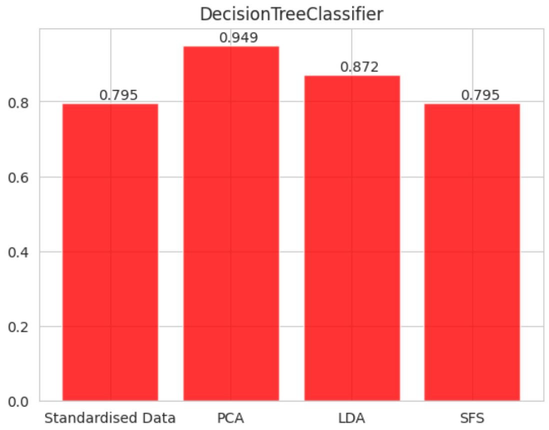


Fig. 17: Plot of accuracy vs. dataset used for training and testing

Here, with decreasing max_depth accuracy, the PCA data increased, which means the PCA data was overfitting. while accuracy remained almost the same for other datasets.

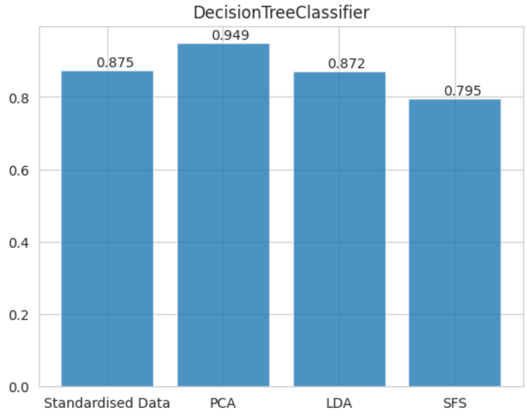


Fig. 18: Plot of F1 score vs. dataset used for training and testing

Implementing Bagging on the Same Decision Tree: (Random Forest)

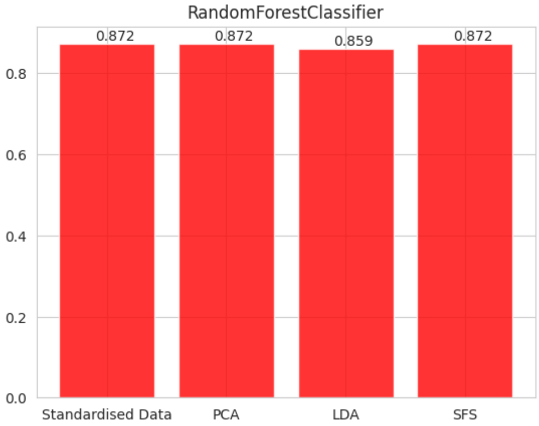


Fig. 19: Plot of accuracy vs. dataset used for training and testing

Comparing the above performance with the one in Fig. 17, we can see that the performance increased on standardised data and SFS, which shows the model was overfitting on those datasets.

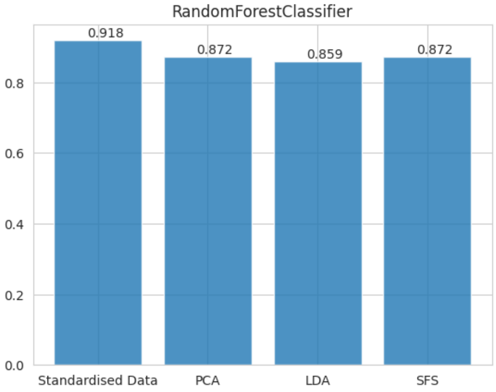


Fig. 20: Plot of F1 score vs. dataset used for training and testing

Implementing Adaboost with Decision Tree as a model:

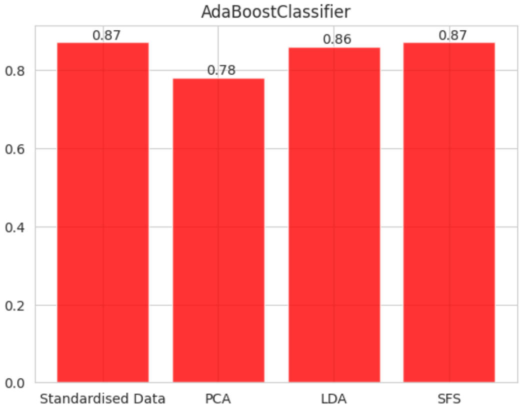


Fig. 21: Plot of accuracy vs. dataset used for training and testing

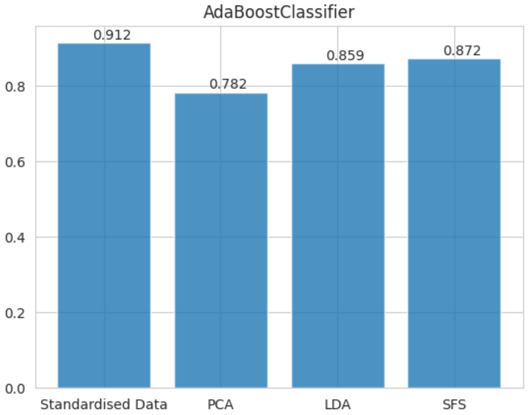


Fig. 22: Plot of F1 score vs. dataset used for training and testing

Implementation of XGBoost:

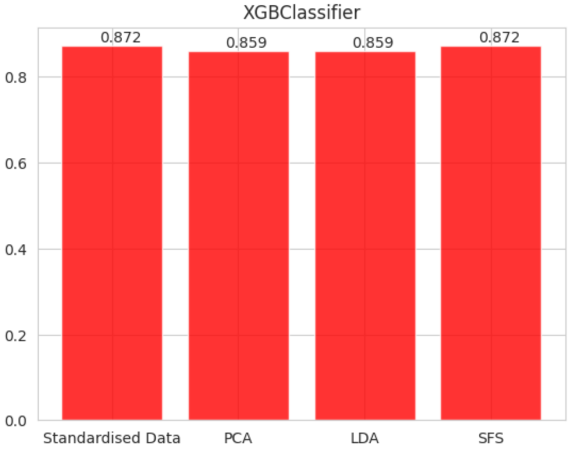


Fig. 23: Plot of accuracy vs. dataset used for training and testing

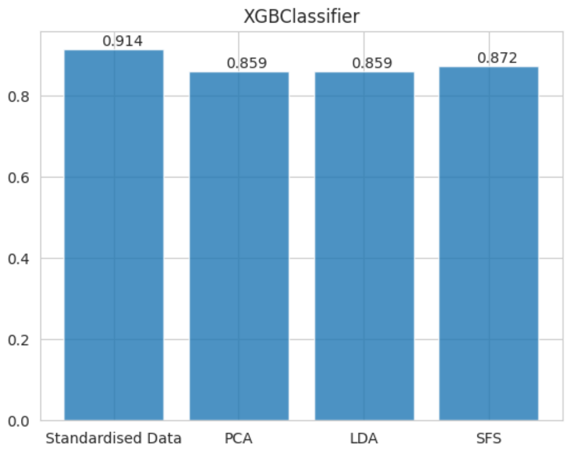


Fig. 24: Plot of F1 score vs. dataset used for training and testing

Neural Network:

First, using Cross validation data, we found the optimal number of hidden layer nodes for highest accuracy, and then we found accuracy and an F1 score on the MLP.

MLP- Relu activation function, 1 hidden layer with 26 nodes.

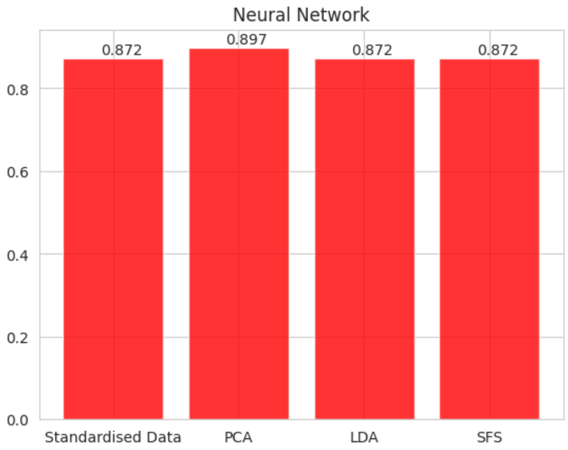


Fig. 25: Plot of accuracy vs. dataset used for training and testing

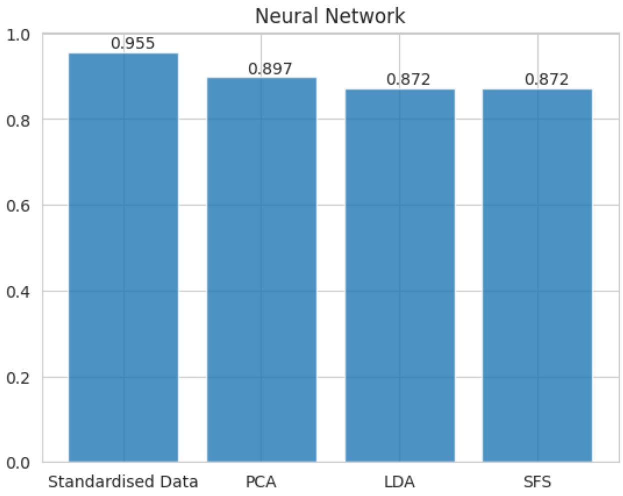


Fig. 26: Plot of F1 score vs. dataset used for training and testing

Support Vector Machines:

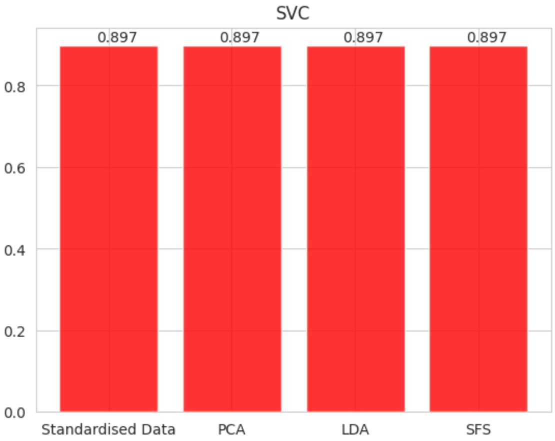


Fig. 27: Plot of accuracy vs. dataset used for training and testing

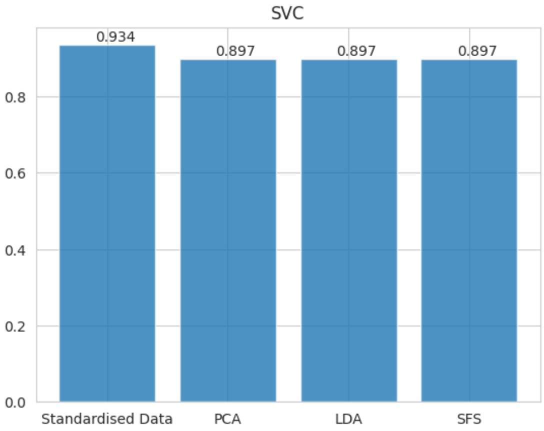


Fig. 28: Plot of F1 score vs. dataset used for training and testing

KNN Classifier:

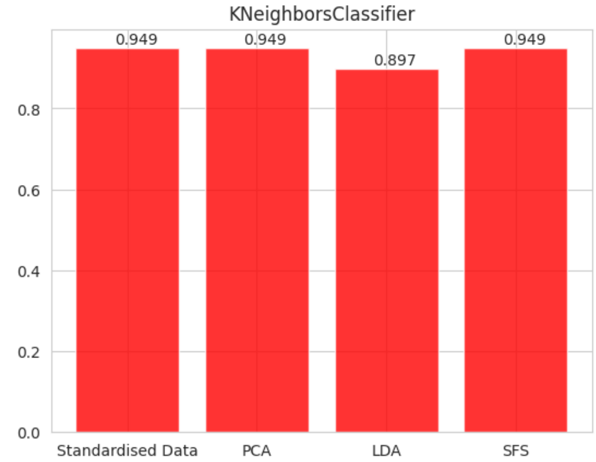


Fig. 29: Plot of accuracy vs. dataset used for training and testing

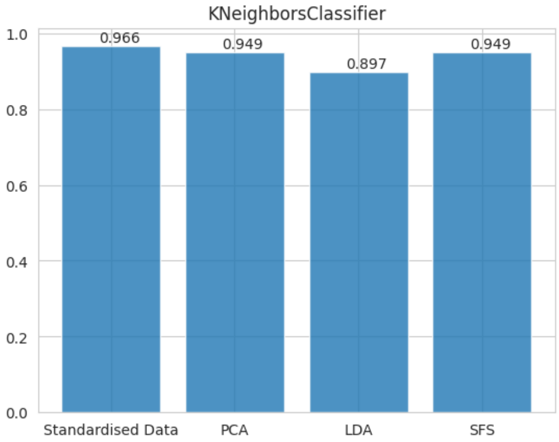
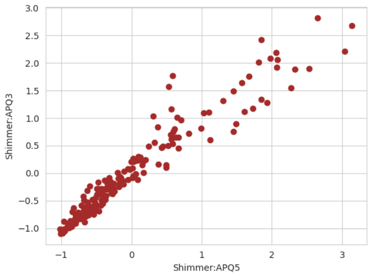


Fig. 30: Plot of F1 score vs. dataset used for training and testing

Out of all the models implemented in this project, KNN gives the best performance with standardised data, as both F1 score and accuracy are at their maximum in that case.

It is also evident from the outlier analysis we did before that the data points are in close proximity to each other; hence, we could have implemented the KNN directly.



But in this project, we were just trying to go from one model to another in a systematic way.

REFERENCES

1) www.wikipedia.com
2) Pattern Classification Second Edition by Duda et. al.
3) Lecture Slides, CSL:2050 Spring Term 2023, Dr. Richa Singh, IIT Jodhpur