Analysis of Electrical Grid Stability data

GROUP 7(THURSDAY BATCH)

Steps to Data Classification Analytics

Data Descriptive Analysis	Data Preprocessing	Classification Model Analytics	Choosing Best Model

About Data

The dataset contains data from the local stability analysis of the 4-node star system (electricity producer is in the center) implementing Decentral Smart Grid Control concept.

tau1 – tau4: The value for electricity producer.

p1 – p4: It gives the nominal power consumed

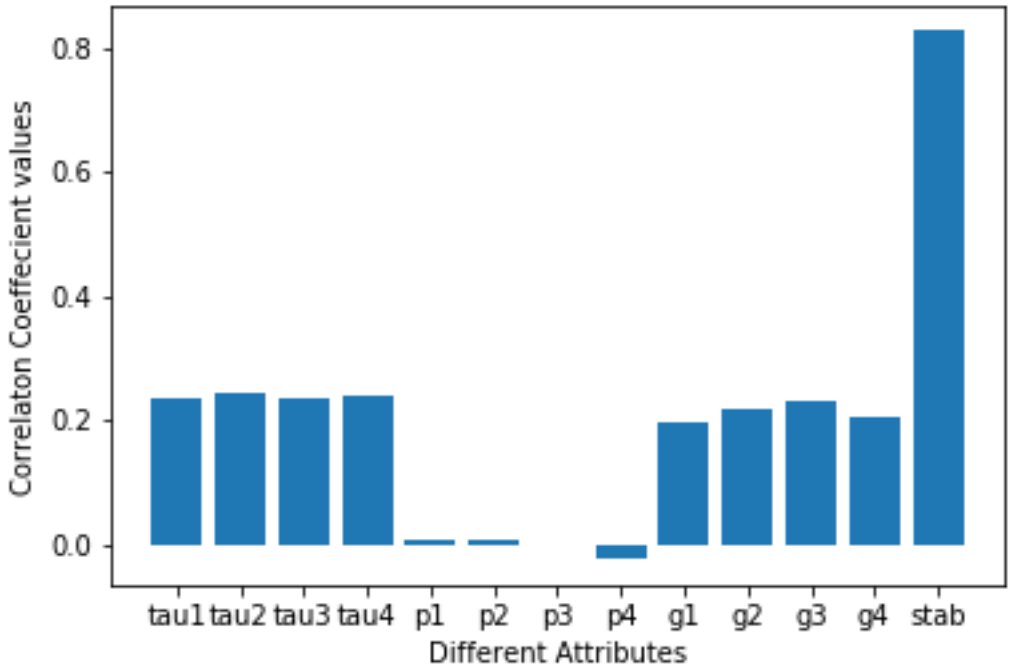
g1 – g4: It gives the coefficient (gamma) proportional to price elasticity

stab: The maximal real part of the characteristic equation root

stabf: the stability label of the system (categorical: stable/unstable).

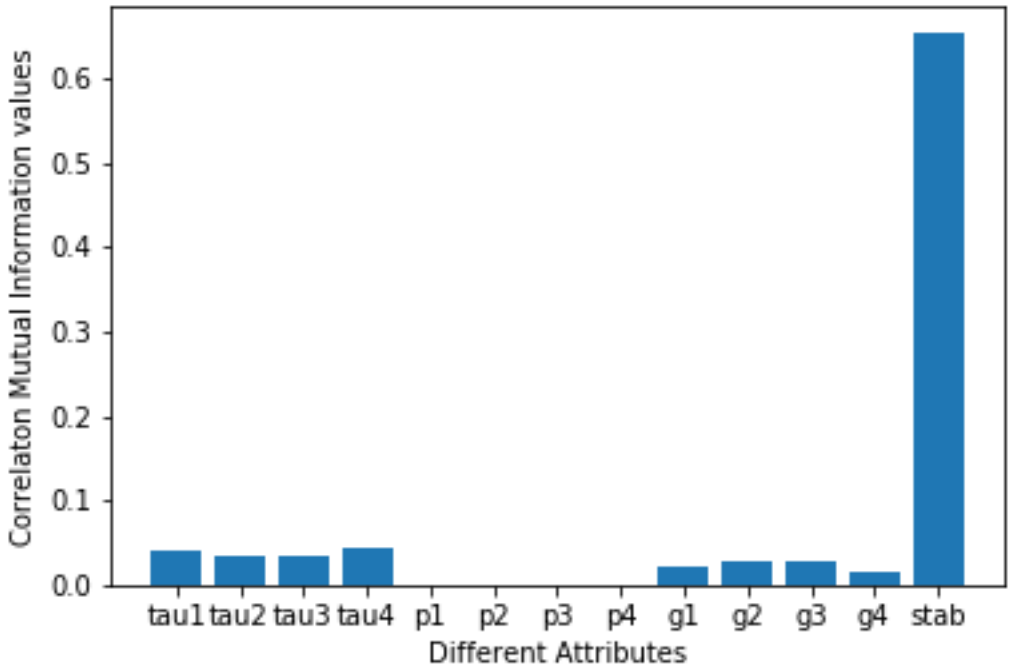
Correlation Analysis of Data

Correlation Coefficient values of different **columns** with target attribute(stabf)



Mutual Information

Mutual information is one of many quantities that measures how much one random variables tells us about another somewhat similar to correlation coefficient but unlike correlation coefficient it also takes into account the possibility of higher order relation among two attributes.



For two discrete variables X and Y whose joint probability distribution is $P_{XY}(x,y)\ ,$ the mutual information between them, denoted $I(X;Y)\ ,$

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log {P_{XY}(x,y) \over P_X(x) P_Y(y)} = E_{P_{XY}} \log{P_{XY} \over P_X P_Y} \, .$$

For Independent Attributes correlation coefficient is 0, but it is not other way round due to possibility of higher order relations

Understanding the Bias-Variance Trade-off

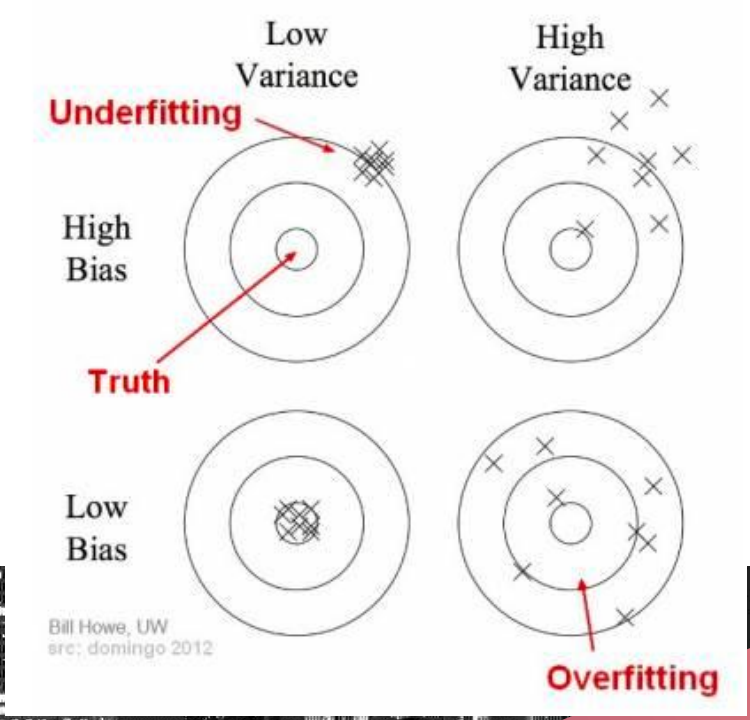
Total Error = Bias^2 + Variance + Irreducible Error

What is Bias?

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

What is Variance?

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.



Total Error

We need to find a good balance between bias and variance such that it minimizes the total error.

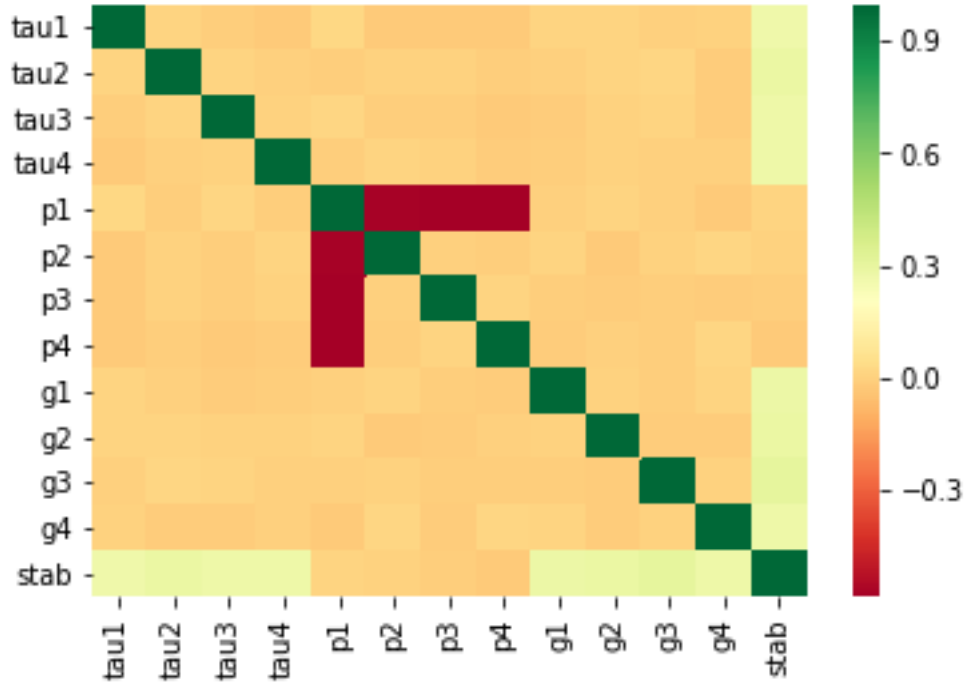
Optimal Balance

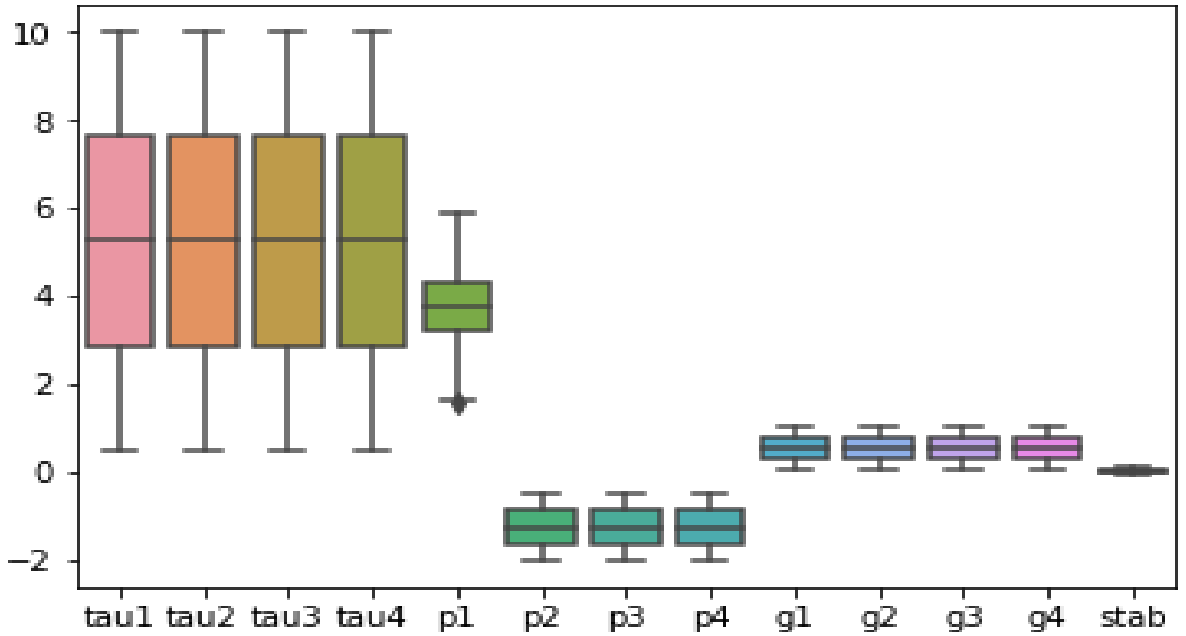
An optimal balance of bias and variance would never overfit or underfit the model.

Data preprocessing

Data Preprocessing:

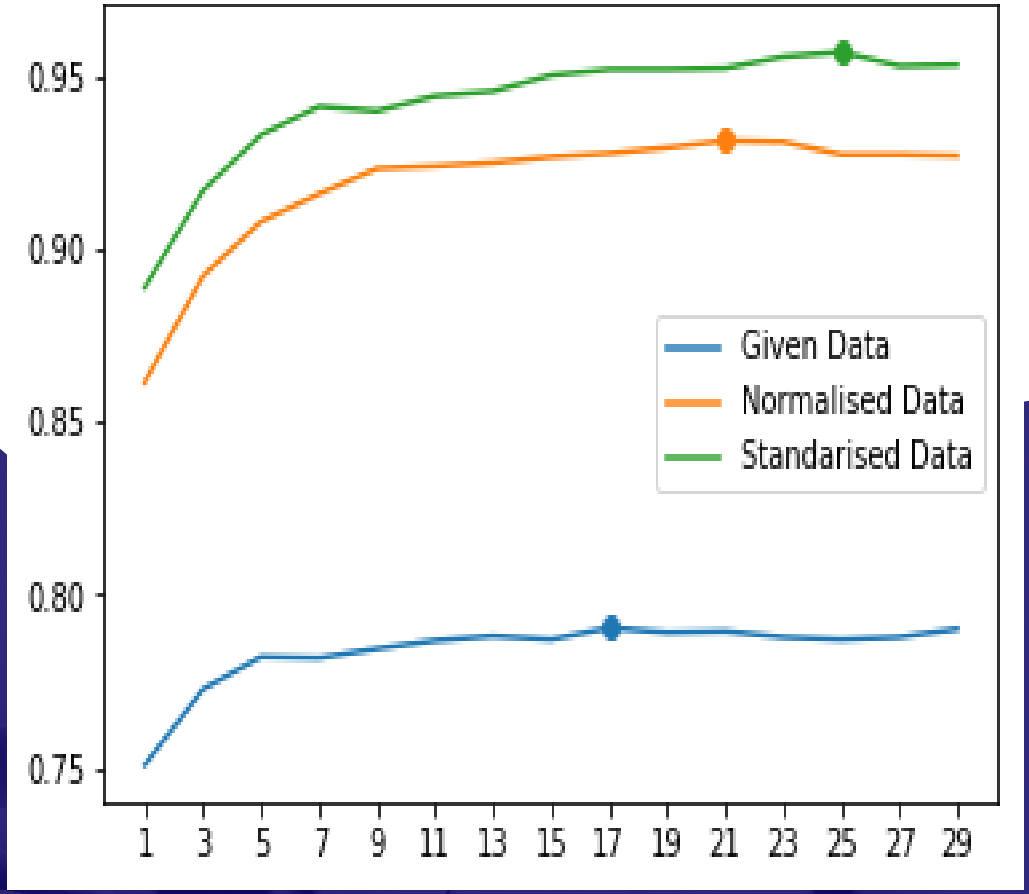
1.No null values in given data.
2.Only one column of data has outliers present.
(we dropped that column P1)
3.We also dropped 2nd last column (because 2nd last column is also target attribute)
 4.For some classification techniques
(i.e. regression) we dropped last column



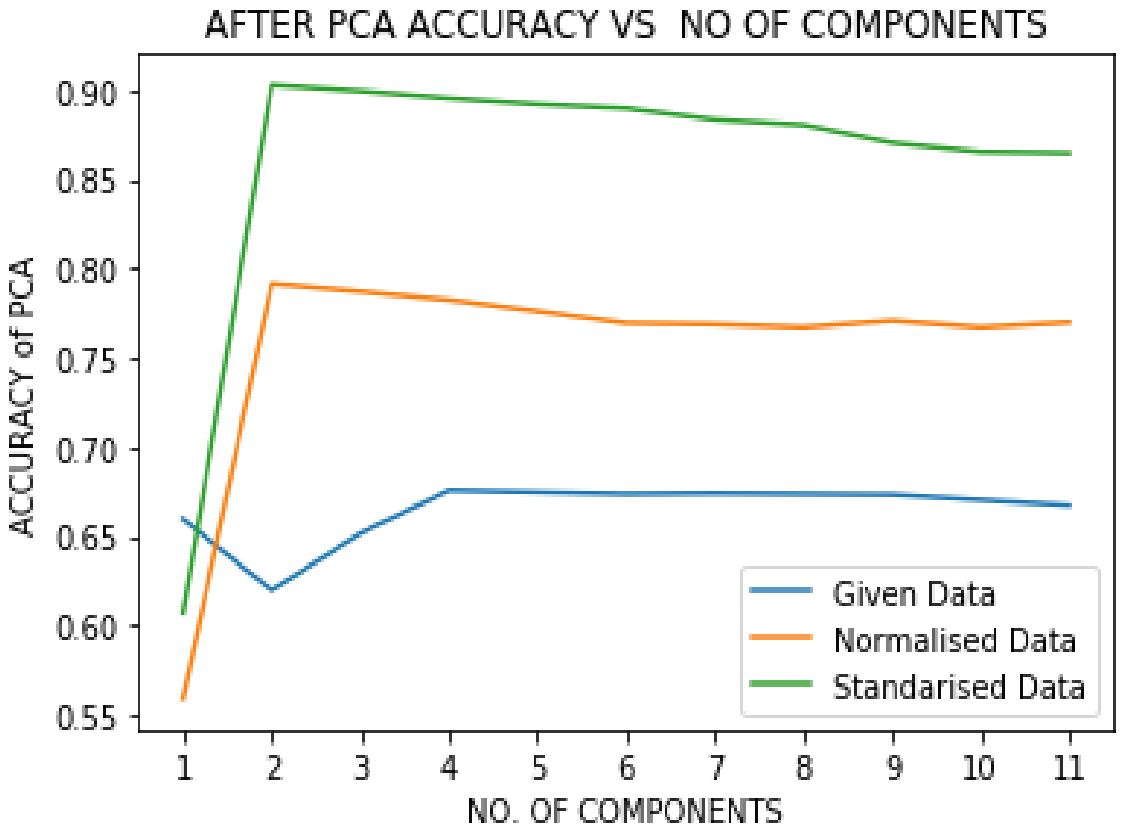


K-nearest neighbour method(without PCA):

For original data: Maximum accuracy on test data = 79% and and train data = 82% for k =17
For normalised data: Maximum accuracy on test data = 93.1% and and train data = 93.8% for k =21
For standardised data: Maximum accuracy on test data = 95.6% and and train data = 95.8% for k =21



K-nearest neighbour method(using PCA):



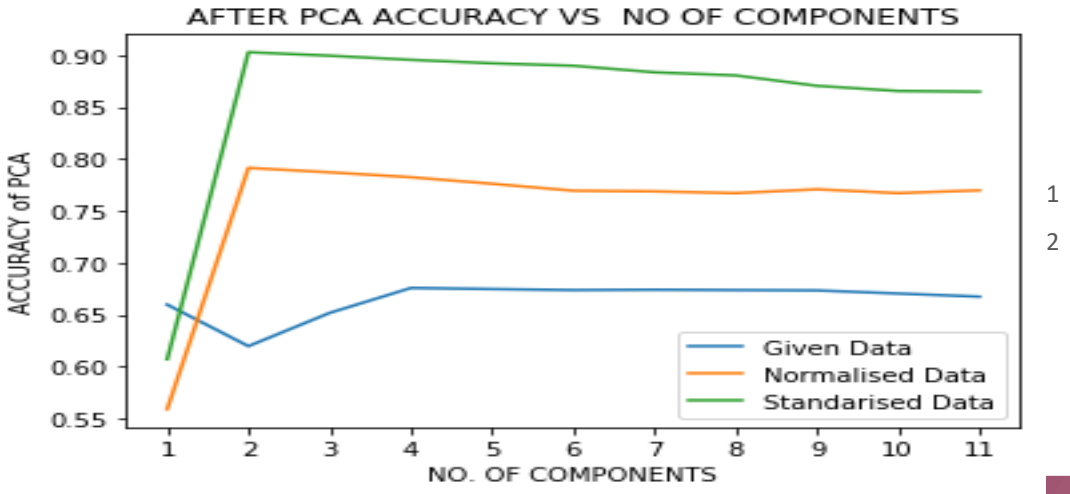
For original data:
Maximum accuracy for test data is 68.1% at n_components= 4

For normalised data:
Maximum accuracy for test data is 78% at n_components= 2

For standardised data:
Maximum accuracy is for test data is 92.2% at n_components= 2

K-nearest neighbour method(using PCA):

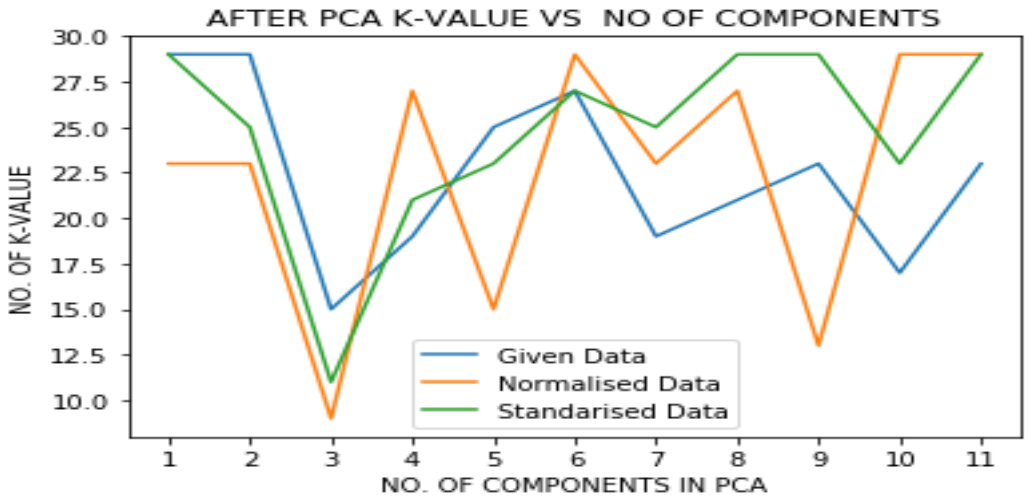
Accuracy Vs Number of components in PCA



1

2

Max k-value VS Number of components in PCA



Inferences taken from graph:

1.k-value(KNN) and accuracy of PCA are uncorrelated.

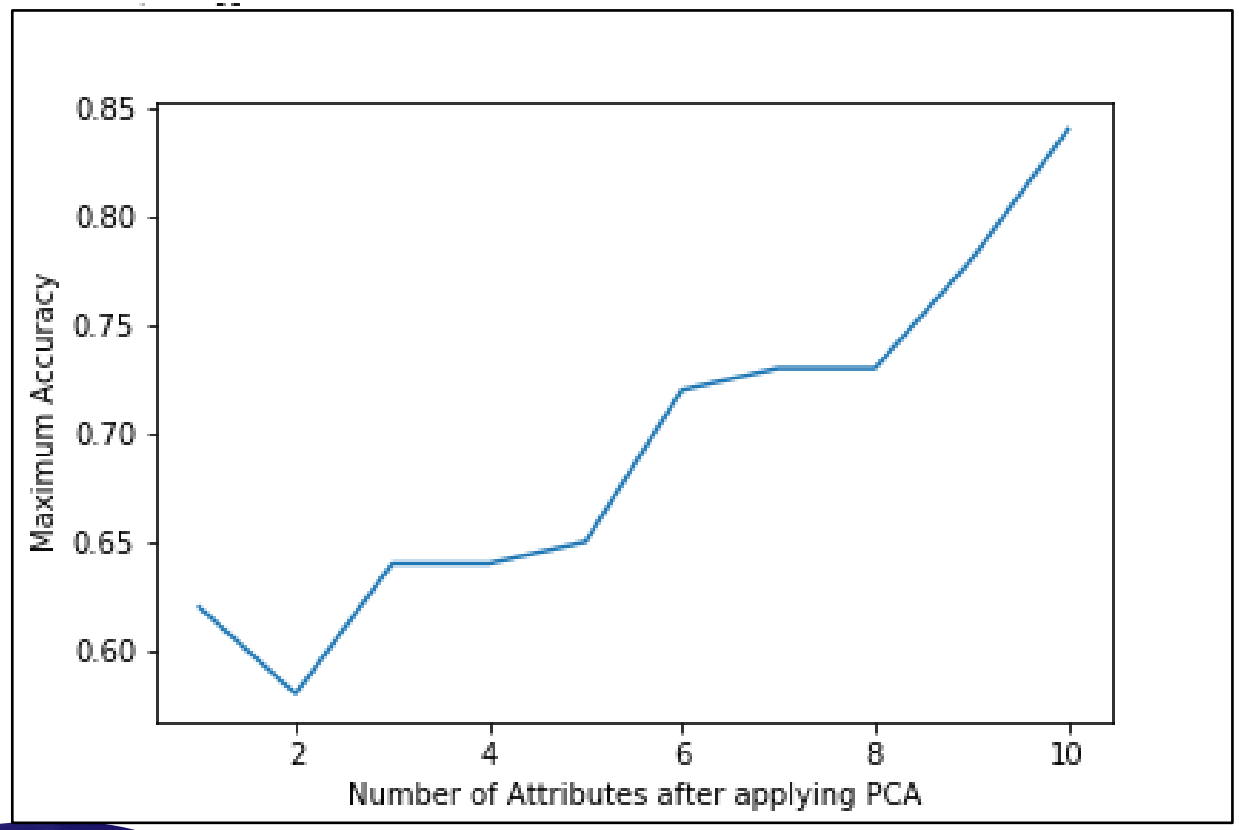
2.K-value(KNN) and number of components of PCA are also uncorrelated.

Gaussian Mixture Model

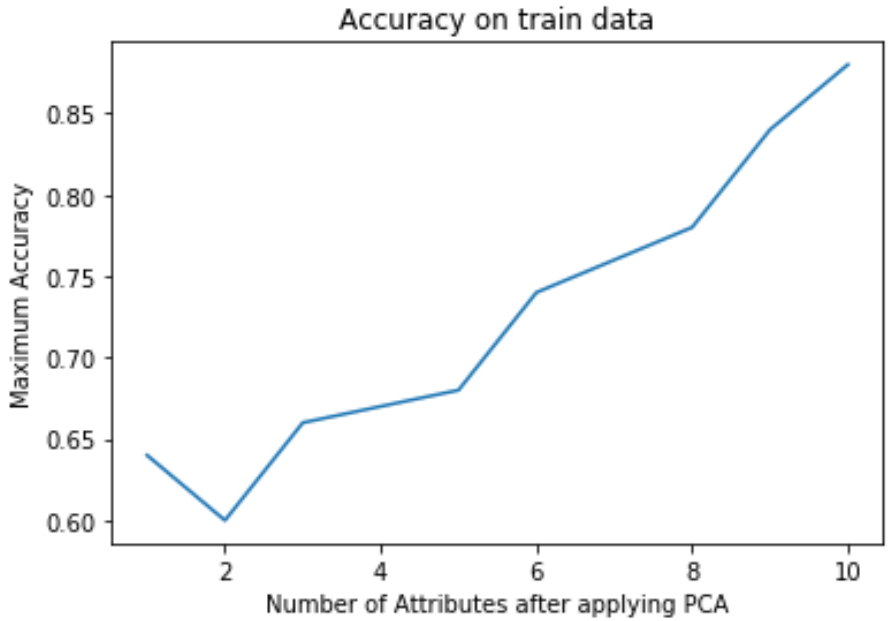
A **Gaussian mixture model** is a probabilistic **model** that assumes all the data points are generated from a **mixture** of a finite number of **Gaussian** distributions with unknown parameters.

From the graph of **Max Acc. Vs No. of attributes** we can see that as our data is lossed by applying PCA, the acc. decreases mostly but with more number of attributes we seem to get **more accurate results**.

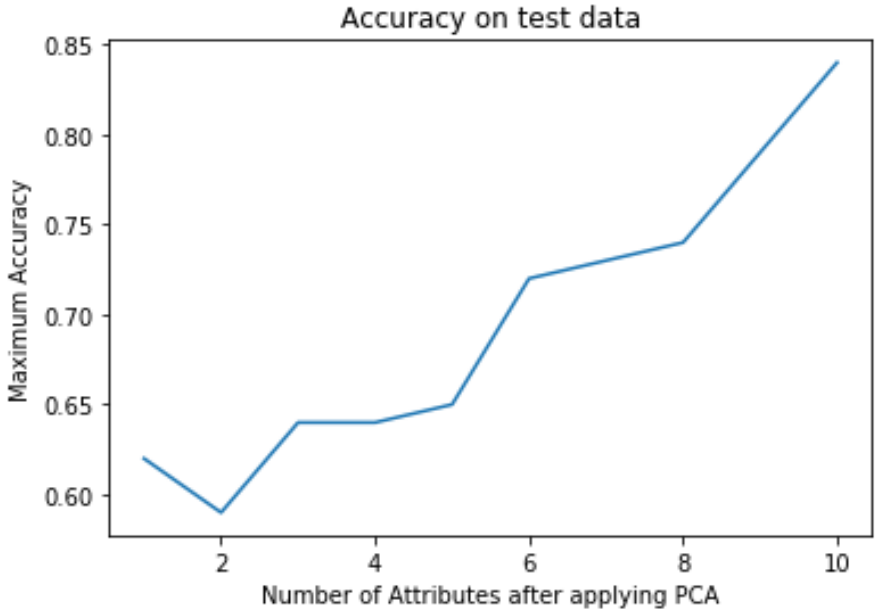
```
score_arr1 = gmm1.score_samples(X_test) * prior1
score_arr2 = gmm2.score_samples(X_test) * prior2
```



Accuracy on Train and Test Data



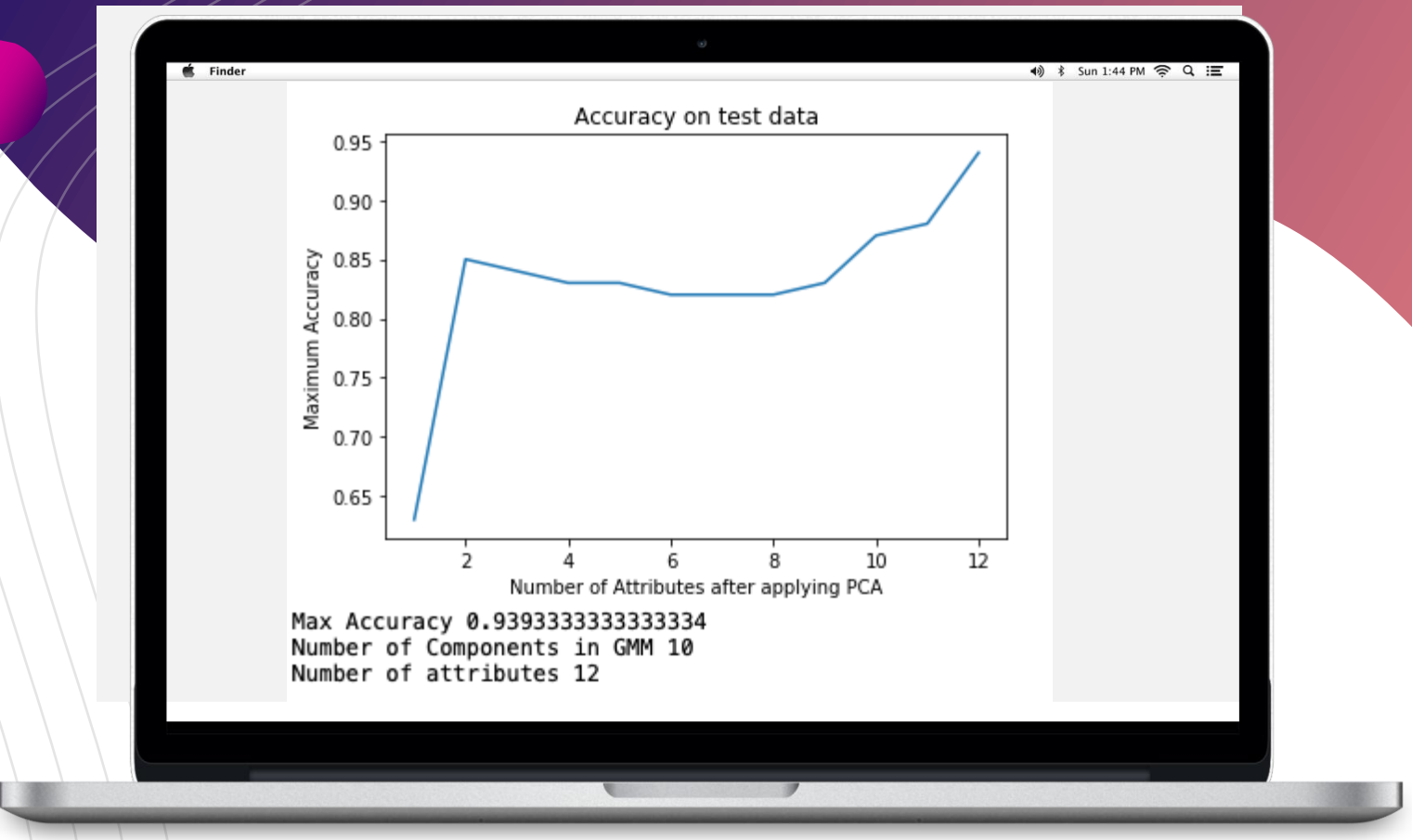

```
Max Accuracy 0.8821428571428571
Number of Components in GMM 10
Number of attributes 10
```




```
Max Accuracy 0.8366666666666667
Number of Components in GMM 10
Number of attributes 10
```

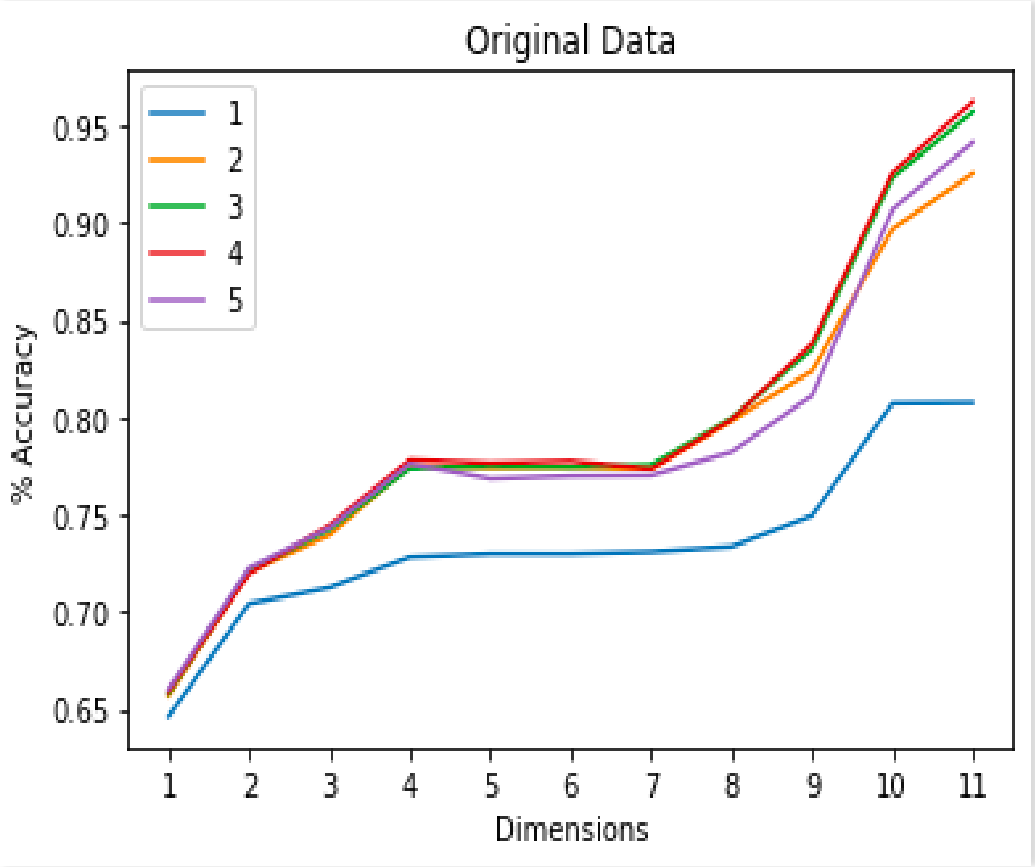
The Max. Accuracy on test data was observed to be about 84% and on train data was observed to be about 89%, this small difference in accuracy of train and test data shows that our model is generalized with low variance and low bias of the model and high accuracy.

Including ‘stab’ column(directly correlated)



On performing GMM, with including another column named ‘stab’ the accuracy increased to 94% as the column was directly correlated with target column.
When the values in this column were positive, the resulting class is stable and unstable for negative values.

Regression (Linear and Polynomial)

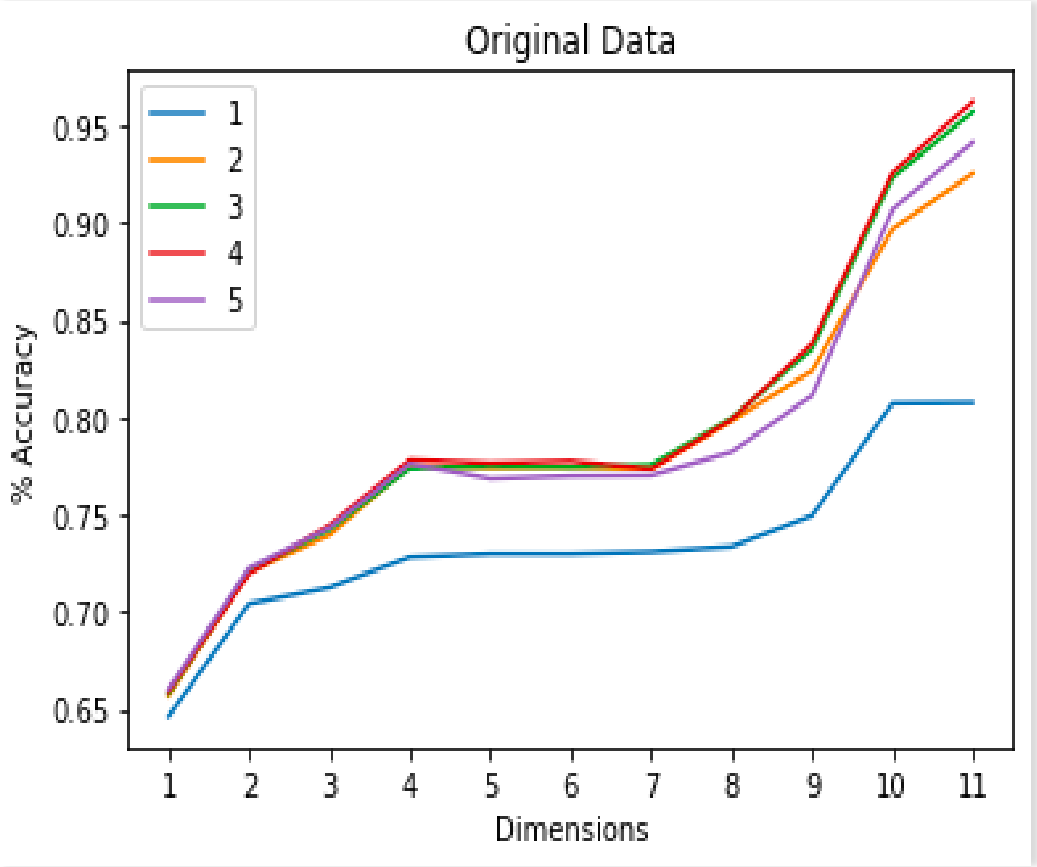


Accuracy considering linear relationship was quite low.

Highest accuracy for degree 4 in even the dimensionally reduced datasets.

Same accuracy for default dimensions in original dataset, standardised dataset and normalised dataset.

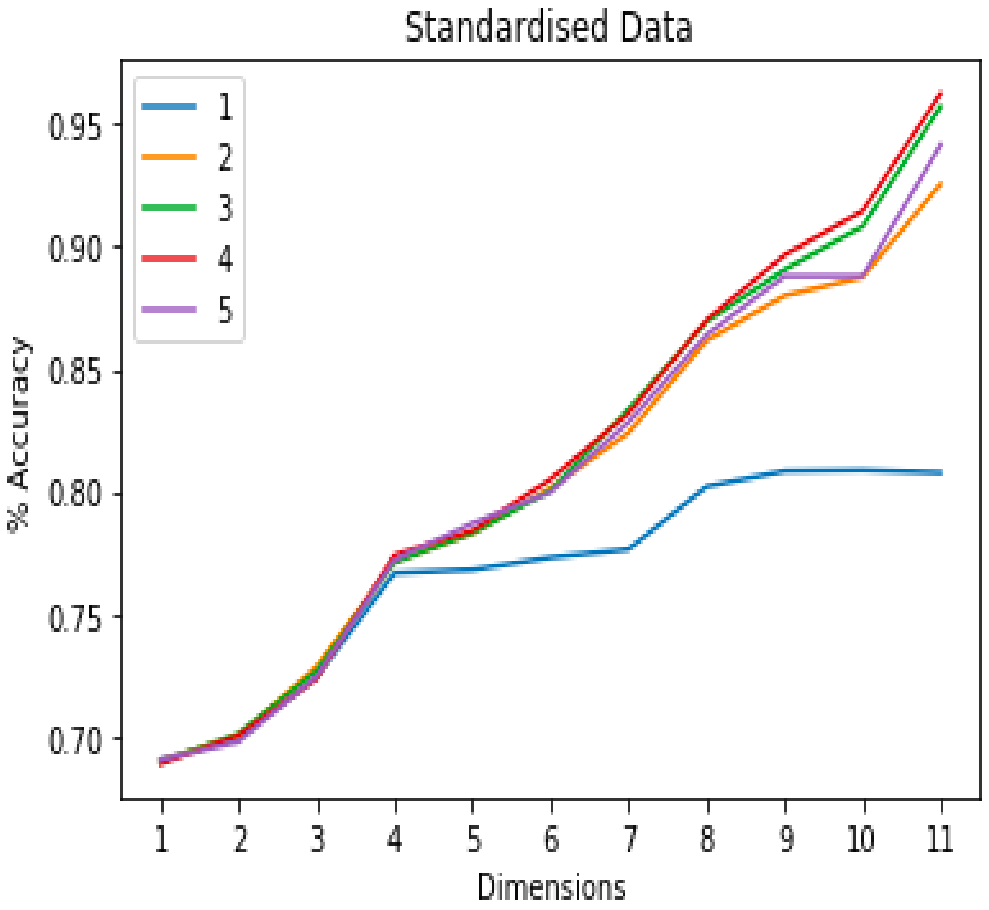
Regression (Linear and Polynomial)

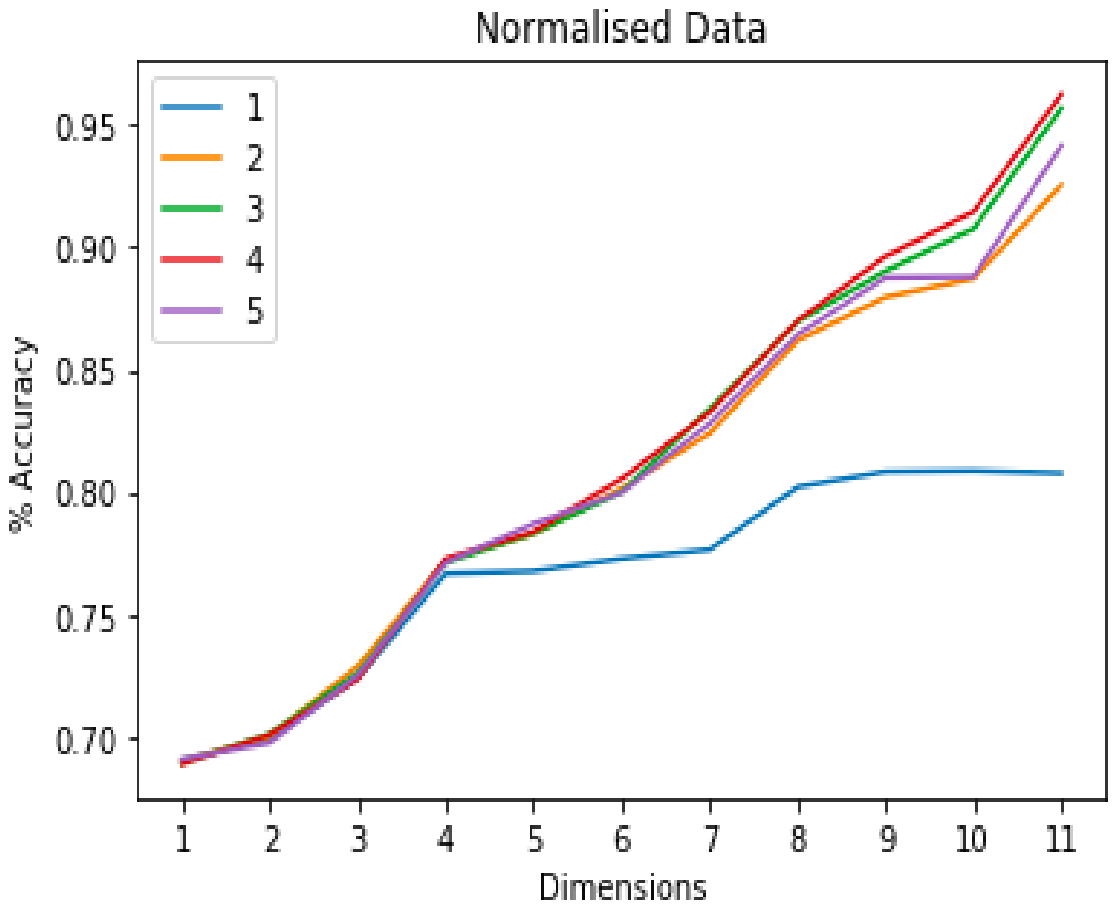


Accuracy for standardised and normalised datasets was more than original dataset.

Both the standardised and the normalised datasets were giving almost the same accuracy but not exactly equal.

Highest accuracy =96.3% (Having original dimensions with polynomial degree four).


Standardised Data
% Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
Dimensions


Normalised Data
% Accuracy
0.95
0.90
0.85
0.80
0.75
0.70
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
Dimensions

Dimensions	Accuracy		
	Original Data	Standardised Data	Normalised Data
1	0.657	0.69	0.69
2	0.72066666666666	0.7	0.7
3	0.73966666666666	0.72833333333333	0.72866666666666
4	0.77533333333333	0.77133333333333	0.772
5	0.77433333333333	0.783	0.783
6	0.77466666666666	0.80133333333333	0.80133333333333
7	0.77366666666666	0.824	0.824
8	0.79833333333333	0.86166666666666	0.862
9	0.824	0.88	0.87966666666666
10	0.897	0.88766666666666	0.88733333333333
11	0.92566666666666	0.92566666666666	0.92566666666666

DATA\PARAMETERS	PRECISION	RECALL	F1-SCORE	SUPPORT
ORIGINAL DATA				
STABLE	0.77	0.71	0.74	2559
UNSTABLE	0.84	0.88	0.86	4441
PCA AND FEATURE SELECTED DATA				
STABLE	0.77	0.71	0.74	2559
UNSTABLE	0.84	0.88	0.86	4441
STANDARDIZED DATA				
STABLE				
UNSTABLE	0.77	0.71	0.74	2559
	0.84	0.88	0.86	4441
NORMALIZED DATA				
STABLE	0.77	0.71	0.74	2559
UNSTABLE	0.84	0.88	0.86	4441

TRAINING PHASE

ACCURACY OF TRAINING PHASE AND CONFUSION MATRIX ENTRIES

- ACCURACY

DATA	ACCURACY
ORIGINAL	0.817857
PCA AND FEATURE REDUCED	0.817857
NORMALIZED	0.817714
STANDARDIZED	0.817428

- CONFUSION MATRIX ENTRIES

DATA\INDEX	00	01	10	11
ORIGINAL	1824	735	540	3901
PCA OR FEATURE REDUCED	1824	735	540	3901
NORMALIZED	1820	739	537	3904
STANDARDIZED	1823	736	542	3899

DATA\PARAMETERS	PRECISION	RECALL	F1-SCORE	SUPPORT
ORIGINAL DATA				
STABLE	0.74	0.70	0.72	1061
UNSTABLE	0.84	0.86	0.85	1939
PCA AND FEATURE SELECTED DATA				
STABLE	0.74	0.70	0.72	1061
UNSTABLE	0.84	0.86	0.85	1939
STANDARDIZED DATA				
STABLE				
UNSTABLE	0.74	0.71	0.72	1061
	0.84	0.86	0.85	1939
NORMALIZED DATA				
STABLE	0.74	0.70	0.72	1061
UNSTABLE	0.84	0.86	0.85	1939

TESTING PHASE

ACCURACY OF TESTING PHASE AND CONFUSION MATRIX ENTRIES

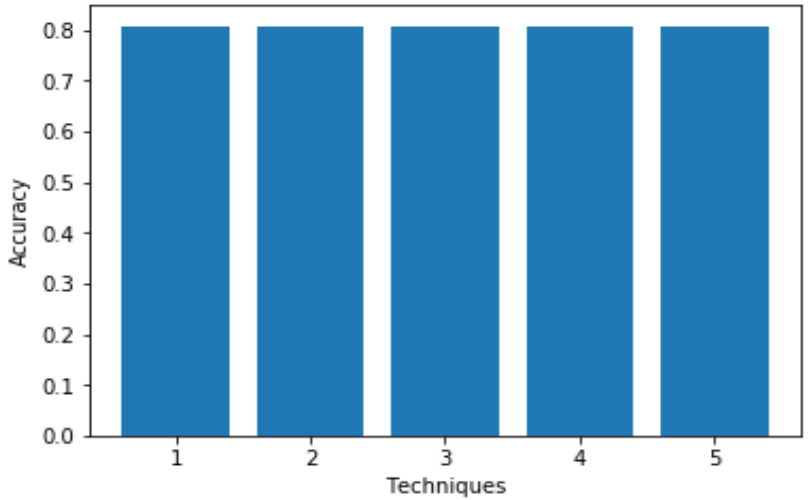
- ACCURACY

DATA	ACCURACY
ORIGINAL	0.807666
PCA AND FEATURE REDUCED	0.807666
NORMALIZED	0.807666
STANDARDIZED	0.807333

- CONFUSION MATRIX ENTRIES

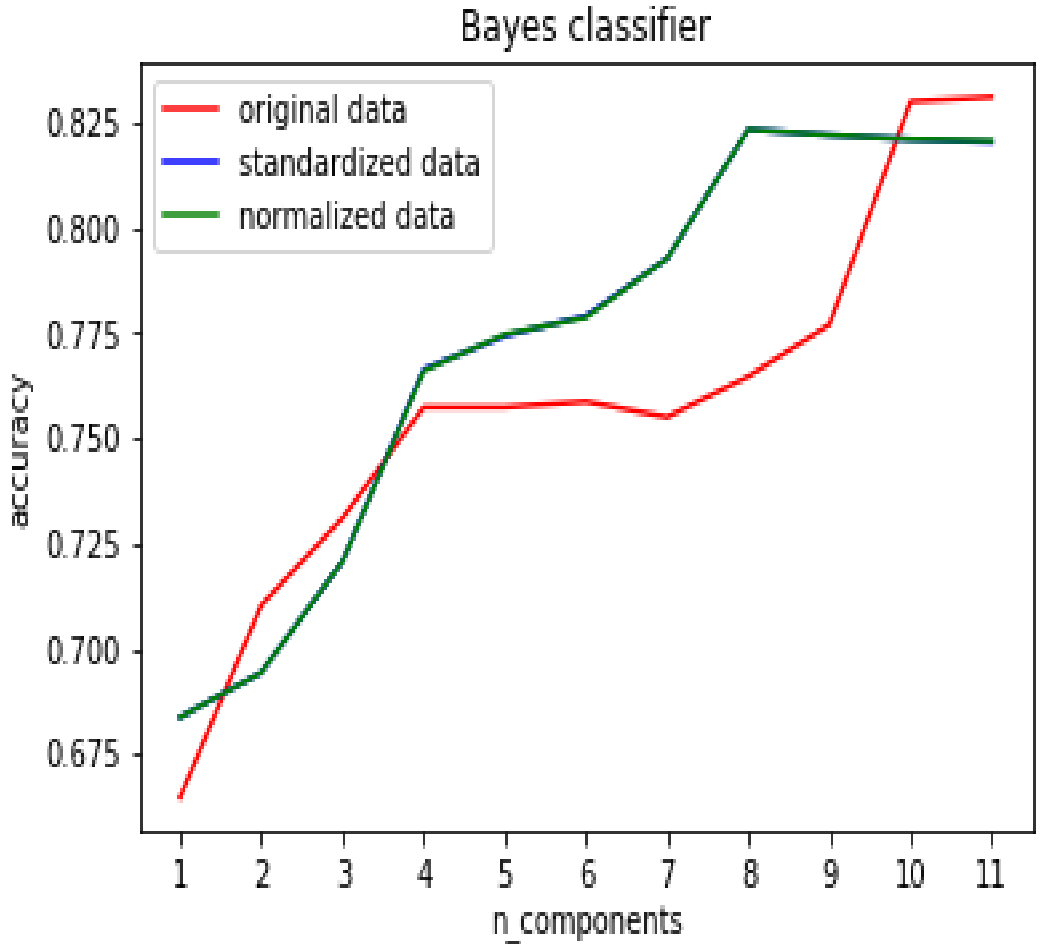
DATA\INDEX	00	01	10	11
ORIGINAL	748	313	264	1675
PCA OR FEATURE REDUCED	748	313	264	1675
NORMALIZED	746	315	262	1677
STANDARDIZED	749	312	266	1673

COMPARISON GRAPHICALLY SUMMARY:



- 1 FOR NORMALIZED DATA.
- 2 FOR STANDARDIZED DATA.
- 3 FOR PCA REDUCED DATA.
- 4 FOR FEATURE SELECTED DATA.
- 5 FOR ORIGINAL DATA.
- *ACCURACY IS OF TESTING PHASE ONLY.*(SAME FOR TRAINING PHASE)
- IN OUR PROJECT PREPROCESSING NOT HELPED A LOT TO ENHANCED THE ACCURACY IN CONTEXT OF LOGISTIC REGRESSION.
- IN GENERAL PREPROCESSING IS VERY IMPORTANT ASPECT FOR ORGANISATIONS WHO DEALS WITH LOT OF BIG DATA.

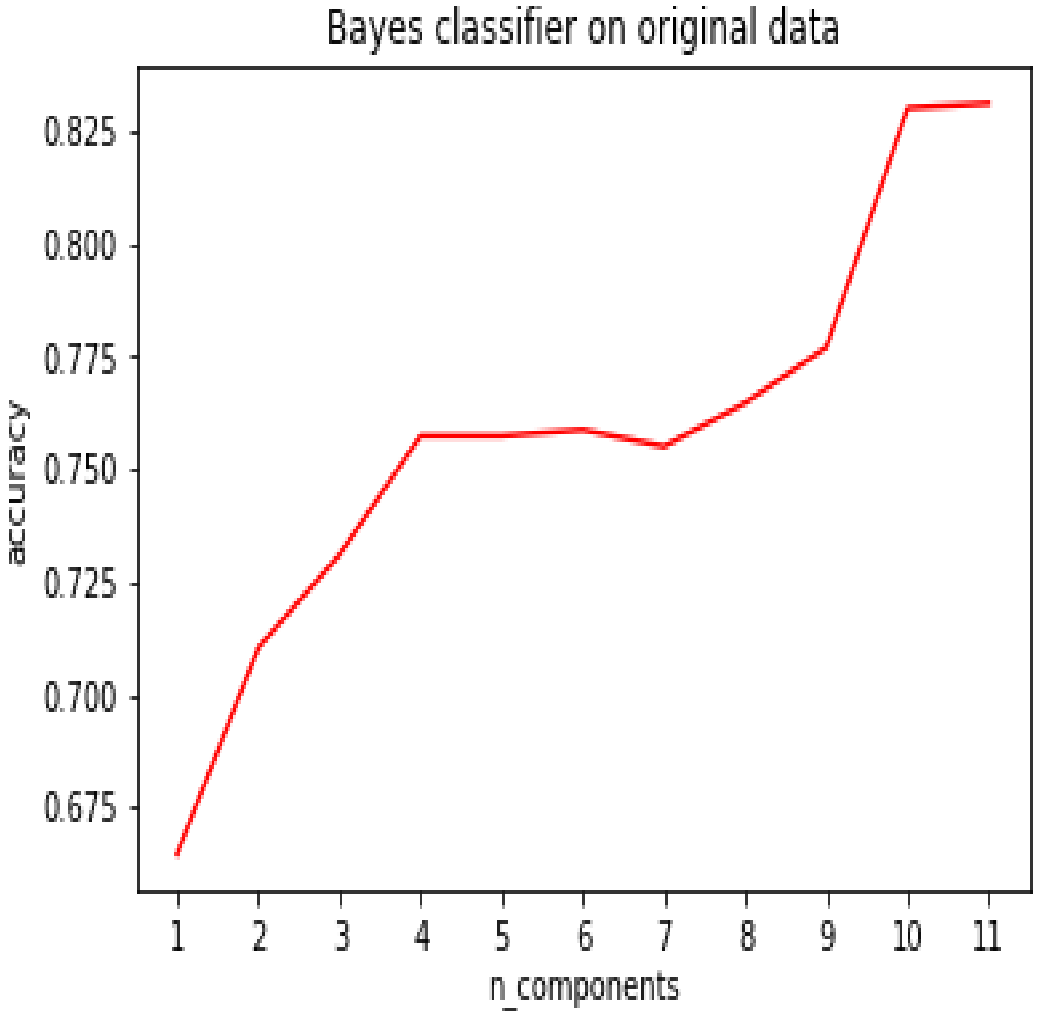
Bayes Classification



Accuracy for original data, standardized data and normalized data without dimensionality reduction is same i.e 83.1%.

For standardized as well as for normalized data accuracy obtained is nearly equal.

Bayes Classification

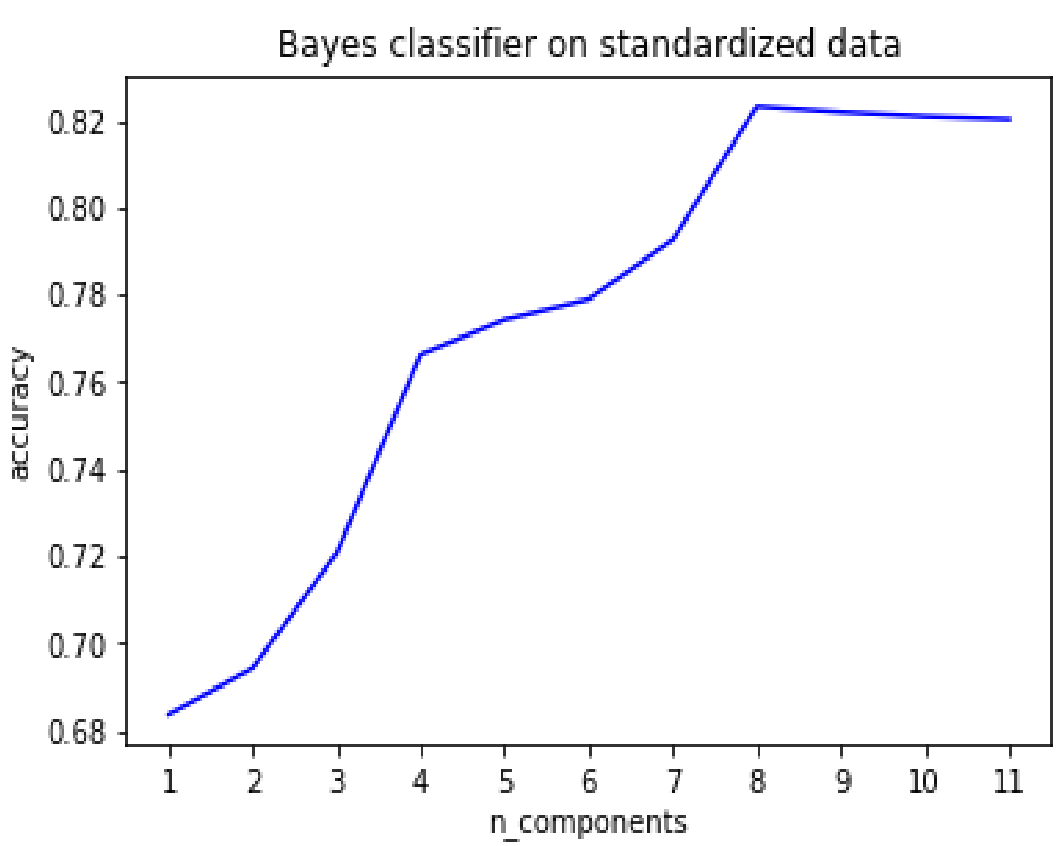


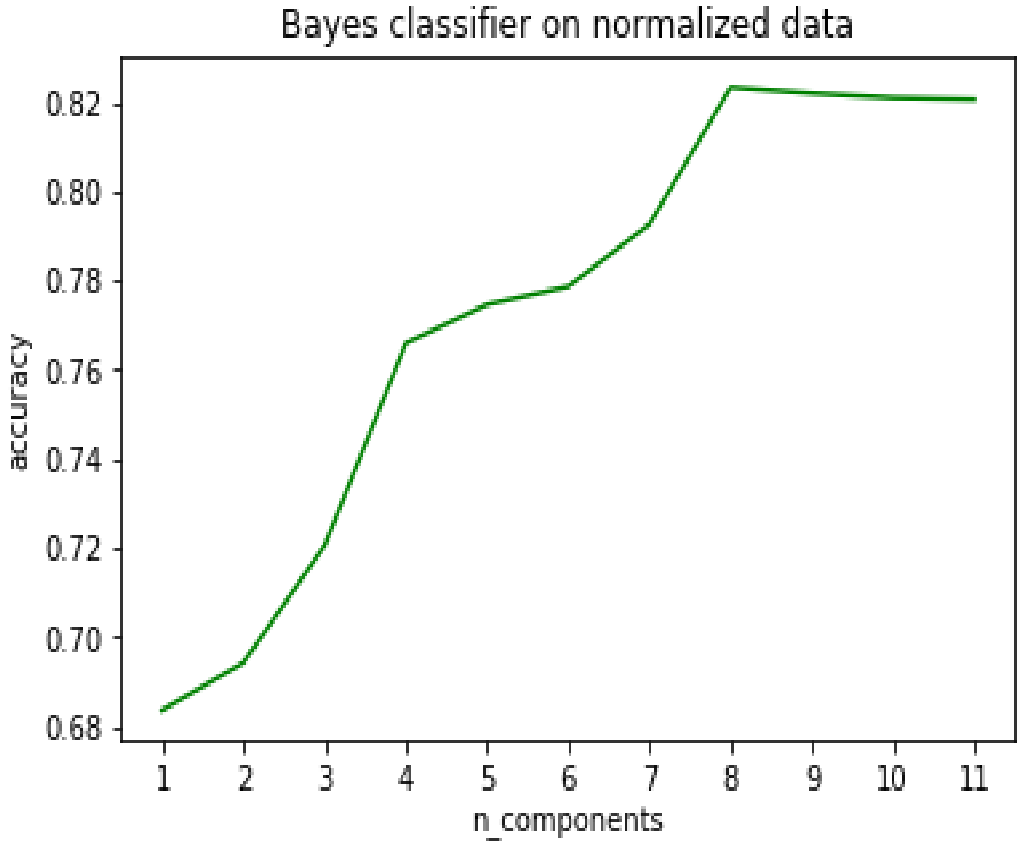
Highest accuracy for original data with reduced dimension is 0.831 for dimension l=11.

Highest accuracy for standardized data with reduced dimension is 0.823 for dimension l=8.

Highest accuracy for normalized data with reduced dimension is 0.823 for dimension l=8.

n_components	Original Data	Standardized Data	Normalized Data
1	0.664666666666	0.6836666666	0.68366666666
2	0.710333333333	0.6943333333	0.69433333333
3	0.731	0.7206666666	0.72066666666
4	0.757333333333	0.7663333333	0.766
5	0.757333333333	0.7743333333	0.77466666666
6	0.758666666666	0.779	0.77866666666
7	0.755	0.7926666666	0.79266666666
8	0.764666666666	0.8233333333	0.82333333333
9	0.777	0.822	0.822
10	0.83	0.821	0.821
11	0.831	0.8203333333	0.82066666666


Bayes classifier on standardized data
accuracy
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
1
2
3
4
5
6
7
8
9
10
11
n_components


Bayes classifier on normalized data
accuracy
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
1
2
3
4
5
6
7
8
9
10
11
n_components

THANK YOU