FEATURE SELECTION TECHNIQUES ON AIR POLLUTION DATASET

Submitted By:
Shreya & Aditi

Content

- INTRODUCTION
- COMMON TERMINOLOGIES
- NORMALIZATION OF DATA
- CORRELATION METHOD EXPLANATION
- PCA METHOD EXPLANATION
- LDA METHOD
- COMPARISON OF ACCURACY RESULTS
- CONCLUSION

INTRODUCTION

Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.

COMMON TERMINOLOGIES

Correlation coefficient: It is a number between -1 and 1 that **tells you the strength and direction of a relationship between variables.** In other words, it reflects how similar the measurements of two or more variables are across a dataset.

Feature Selection Algorithms:

There are three type of feature selection we used in our dataset-

- PCA-The Principal Component Analysis is **a popular unsupervised learning technique for reducing the dimensionality of data**
- LDA-Linear Discriminant Analysis (LDA) is one of the commonly used dimensionality reduction techniques in machine learning to solve more than two-class classification problems.
- Correlation Method-Correlation is **a statistical calculation that indicates that two variables are parallelly related** (which means that the variables change together at a constant rate).

Difference between PCA and LDA

PCA is an unsupervised learning algorithm while LDA is a supervised learning algorithm. This means that PCA finds directions of maximum variance regardless of class labels while LDA finds directions of maximum class separability.

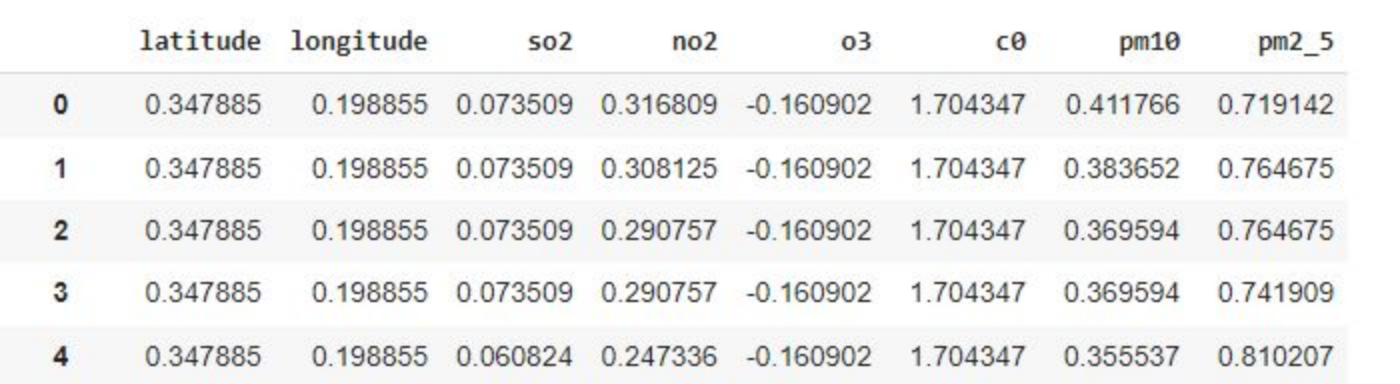
Normalization:

Normalization generally refers to **processes that achieve scales between zero and one**.

Standardization:

It uses a principle called the standard deviation to describe the distribution of the data points.

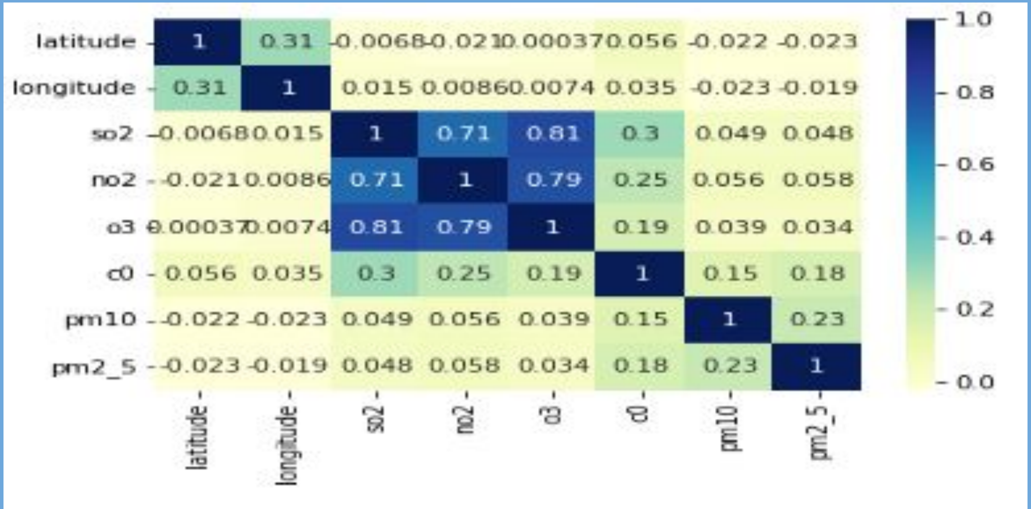
NORMALIZATION OF DATA



	latitude	longitude	so2	no2	o3	c0	pm10	pm2_5
0	0.347885	0.198855	0.073509	0.316809	-0.160902	1.704347	0.411766	0.719142
1	0.347885	0.198855	0.073509	0.308125	-0.160902	1.704347	0.383652	0.764675
2	0.347885	0.198855	0.073509	0.290757	-0.160902	1.704347	0.369594	0.764675
3	0.347885	0.198855	0.073509	0.290757	-0.160902	1.704347	0.369594	0.741909
4	0.347885	0.198855	0.060824	0.247336	-0.160902	1.704347	0.355537	0.810207

METHOD1: CORRELATION METHOD

Correlation values of our normalized dataset through Heatmap.

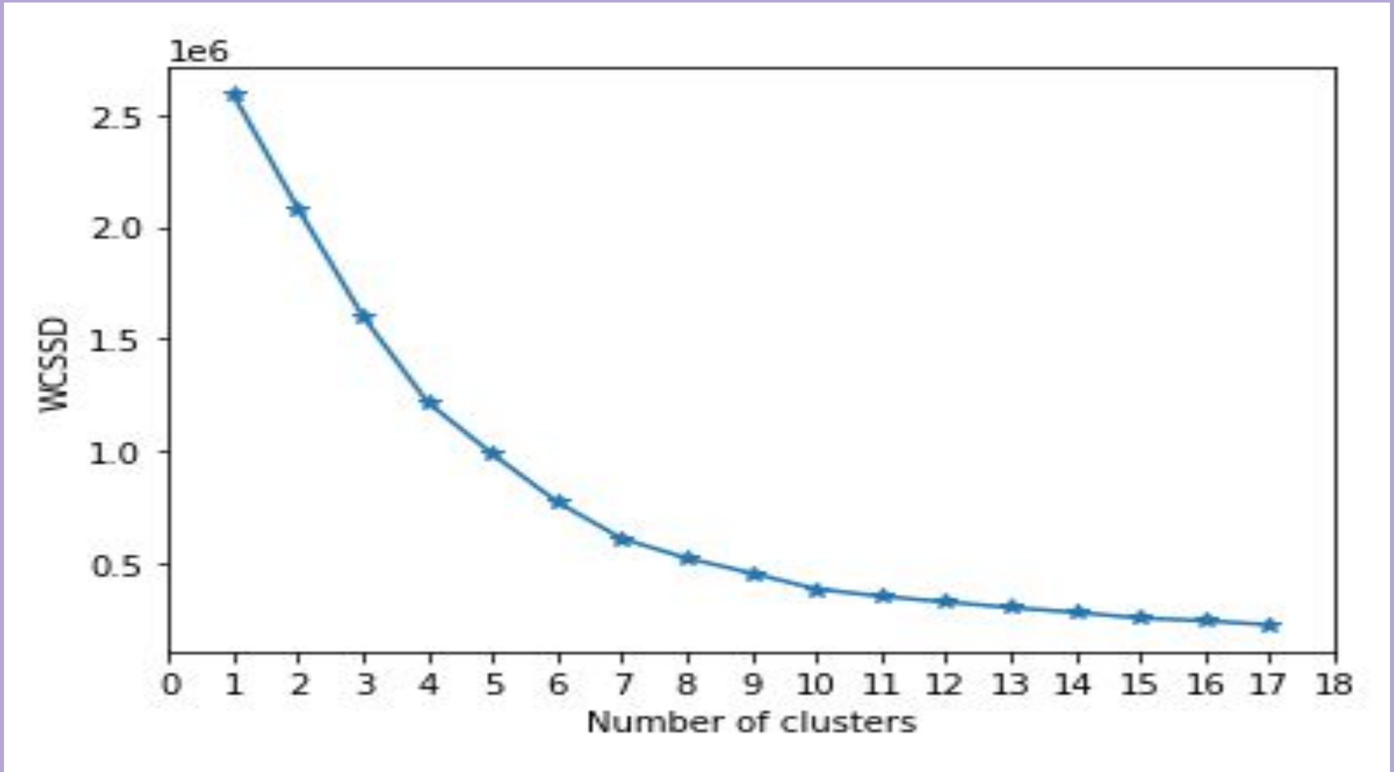


	latitude	longitude	so2	no2	o3	c0	pm10	pm2_5
latitude	1	0.31	-0.0068	-0.021	0.00037	0.056	-0.022	-0.023
longitude	0.31	1	0.015	0.0086	0.0074	0.035	-0.023	-0.019
so2	-0.0068	0.015	1	0.71	0.81	0.3	0.049	0.048
no2	-0.021	0.0086	0.71	1	0.79	0.25	0.056	0.058
o3	0.00037	0.0074	0.81	0.79	1	0.19	0.039	0.034
c0	0.056	0.035	0.3	0.25	0.19	1	0.15	0.18
pm10	-0.022	-0.023	0.049	0.056	0.039	0.15	1	0.23
pm2_5	-0.023	-0.019	0.048	0.058	0.034	0.18	0.23	1

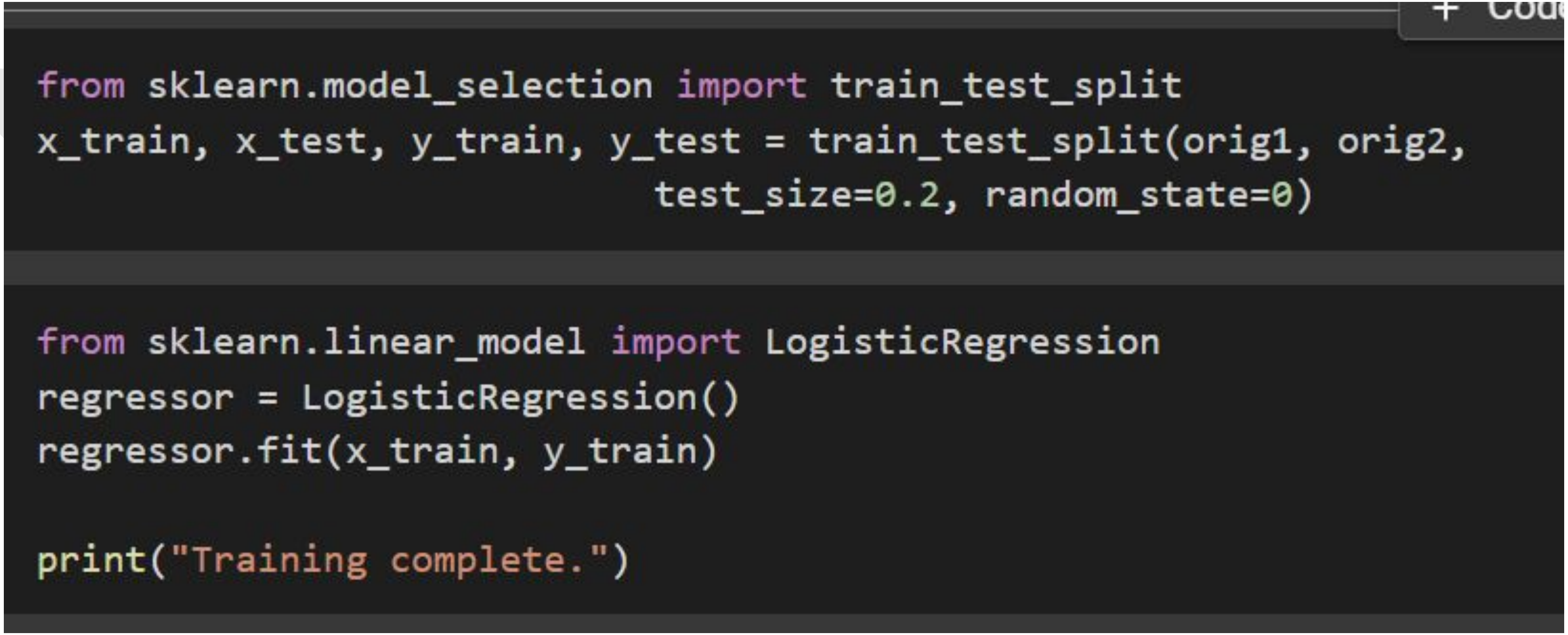
Explanation of Heatmap

As we can see from above heatmap and correlation table the correlation between O3 and sO2 ,O3 and NO2 are high so we will take only one feature from them i.e O3 and drop other two feature i.e. NO2, SO2 and we also drop latitude and longitude columns as they do not give any useful information.

In order to find accuracy of this method we have to label our extracted dataset by applying clustering algorithm i.e. K-Means algorithm.



Optimal number of cluster comes out to be 10 in this case, so now K-means would be applied for 10 clusters.



```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(orig1, orig2,
                            test_size=0.2, random_state=0)
```

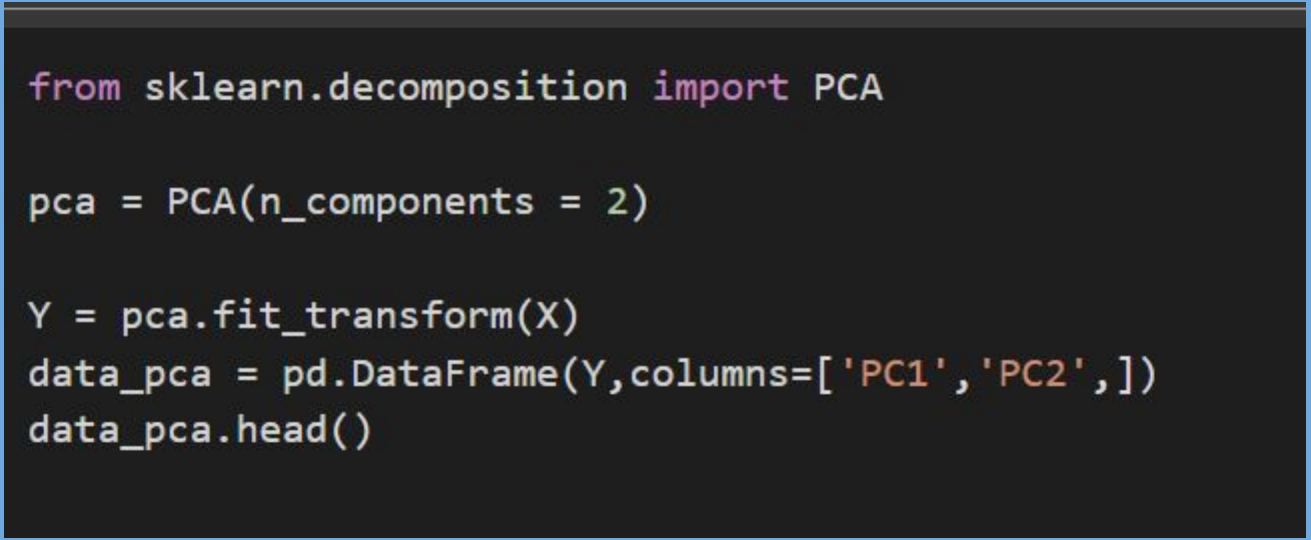
```
from sklearn.linear_model import LogisticRegression
regressor = LogisticRegression()
regressor.fit(x_train, y_train)

print("Training complete.")
```

We are fitting Logistic regression on our data for prediction in above snippet.

METHOD 2: PCA METHOD

Below we are applying PCA for reducing from 8 dimensions to 2 dimensions.

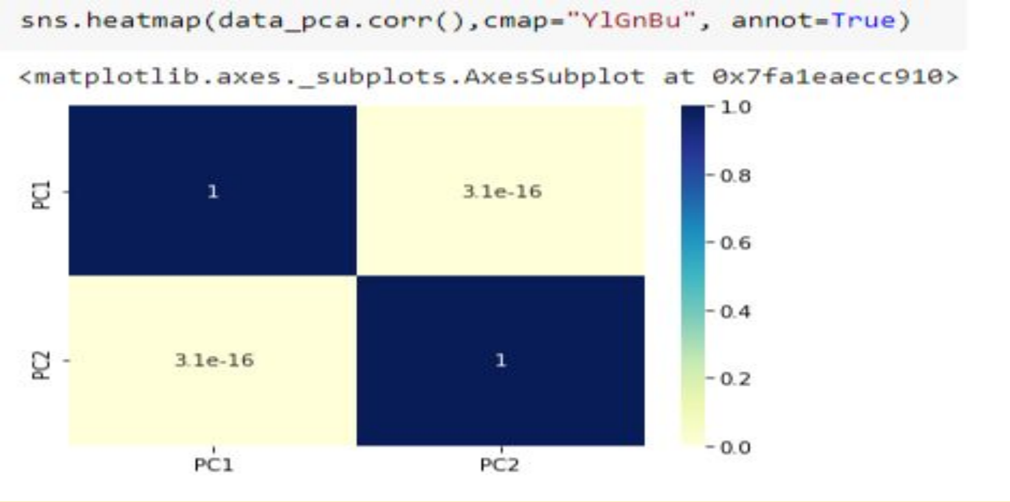


```
from sklearn.decomposition import PCA

pca = PCA(n_components = 2)

Y = pca.fit_transform(X)
data_pca = pd.DataFrame(Y,columns=['PC1','PC2',])
data_pca.head()
```

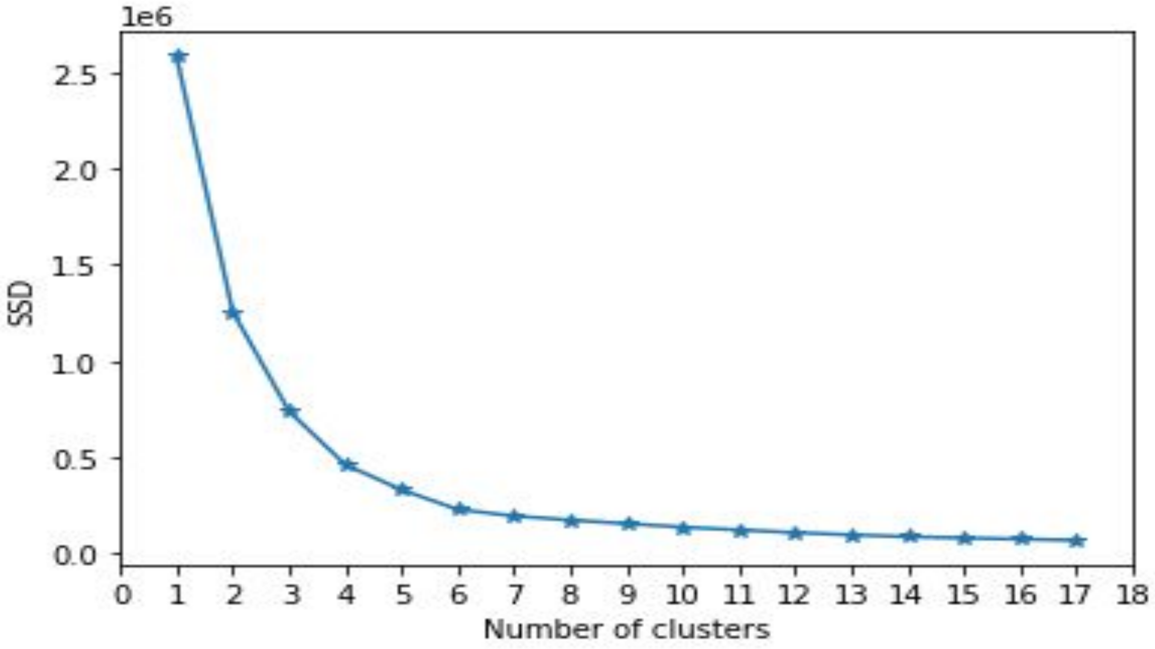
We took normalized data and applied PCA algorithm for dimensionality reduction, took top 2 columns which cover most of our data with less correlation between them as shown in heatmap.



```
sns.heatmap(data_pca.corr(),cmap="YlGnBu", annot=True)
```

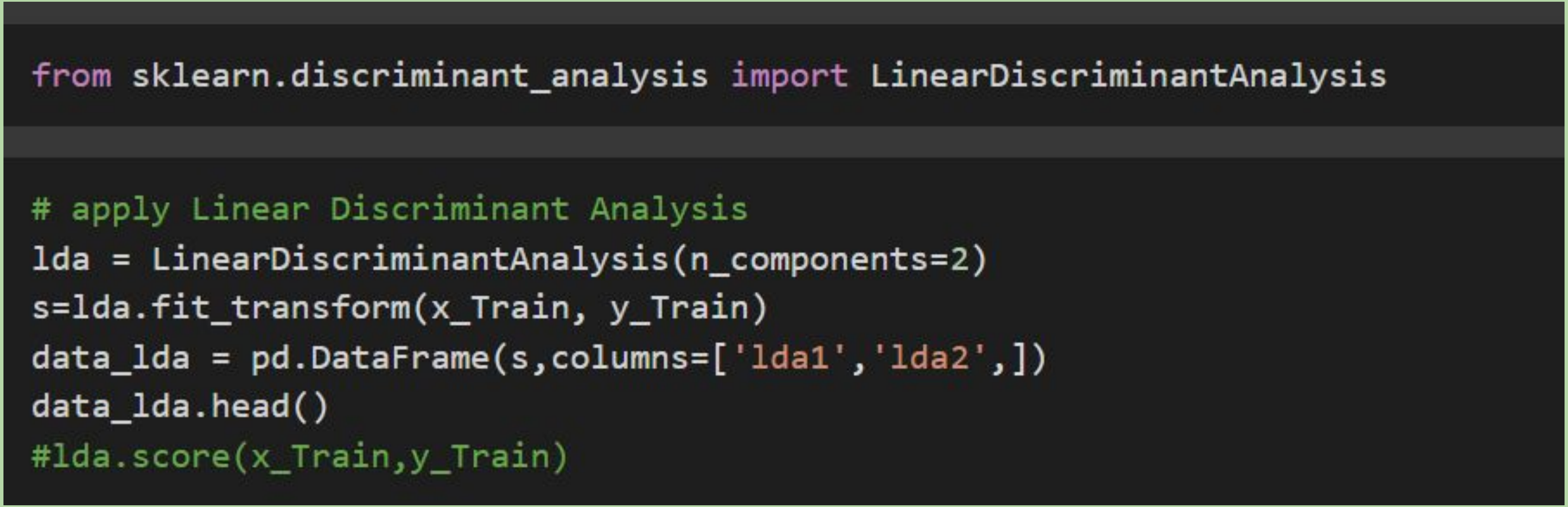
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa1eaecc910>
```

Now for predicting the accuracy of pca firstly ,we have to label our dataset by doing clustering. For this we have to apply the K-means algorithm. Optimal number of cluster comes out to be 6 through WCSS for K-means.



METHOD 3: LDA METHOD

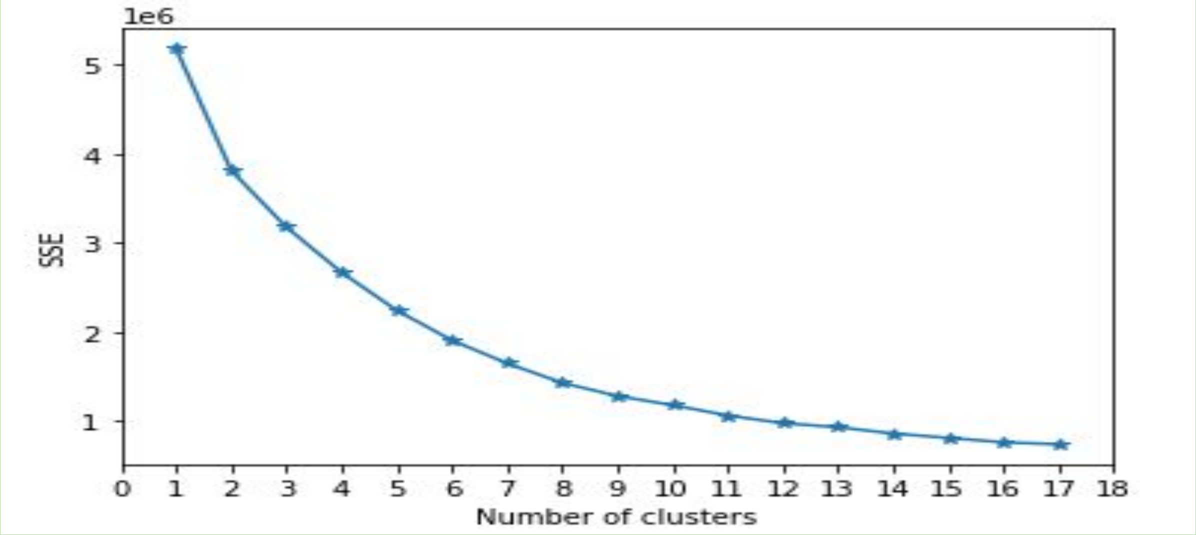
Below we are applying LDA and changing 8 dimensional data to 2 dimensional data.



```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
```

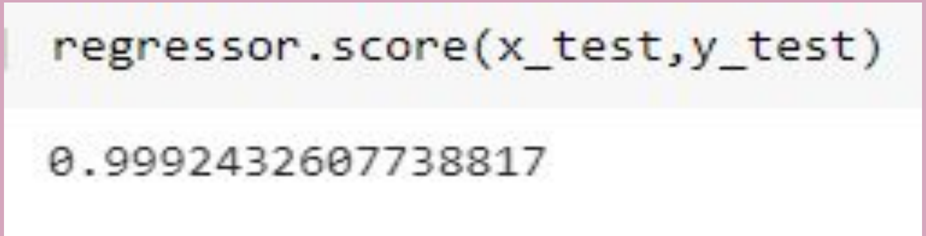
```
# apply Linear Discriminant Analysis
lda = LinearDiscriminantAnalysis(n_components=2)
s=lda.fit_transform(x_Train, y_Train)
data_lda = pd.DataFrame(s,columns=['lda1','lda2',])
data_lda.head()
#lda.score(x_Train,y_Train)
```

The LDA algorithm is applied on a labeled dataset so for making our unlabeled dataset to labeled dataset ,we have to do clustering. For this we have to apply the K-means algorithm. Optimal number of clusters comes out to be 16 through WCSS for K-means clustering.



ACCURACY COMPARISON OF ALL ALGORITHMS

Correlation method



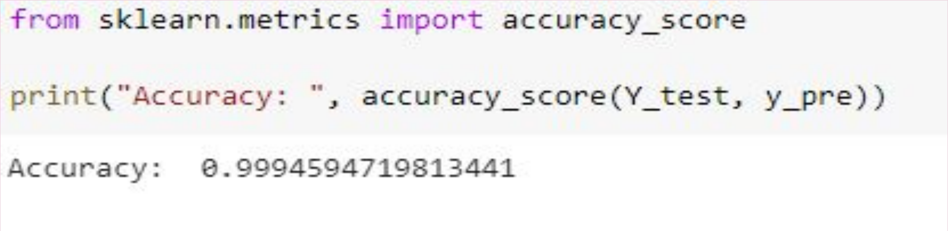
```
regressor.score(x_test,y_test)
```

```
0.9992432607738817
```

Above snippet shows accuracy score of Correlation method comes out to be 99.92%

PCA

After fitting logistic regression model in our dataset the accuracy of our method comes out to be 99.94% as shown in the below snippet.



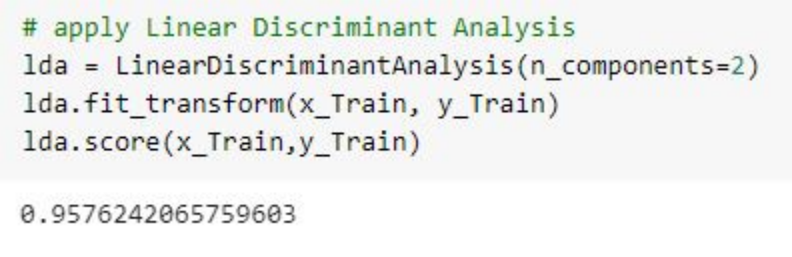
```
from sklearn.metrics import accuracy_score

print("Accuracy: ", accuracy_score(Y_test, y_pre))
```

```
Accuracy:  0.9994594719813441
```

LDA

After fitting Linear discriminant analysis model in our dataset the accuracy of our method comes out to be 95.76% as shown in the below snippet.



```
# apply Linear Discriminant Analysis
lda = LinearDiscriminantAnalysis(n_components=2)
lda.fit_transform(x_Train, y_Train)
lda.score(x_Train,y_Train)
```

```
0.9576242065759603
```

CONCLUSION

From the above three methods we found that Principal component analysis algorithm is best suited for feature selection or dimensionality reduction technique on the given dataset, PCA’s accuracy is 99.94% on the other hand the accuracy score for LDA algorithm and correlation method are 95.76% and 99.24% respectively.

THANK YOU!