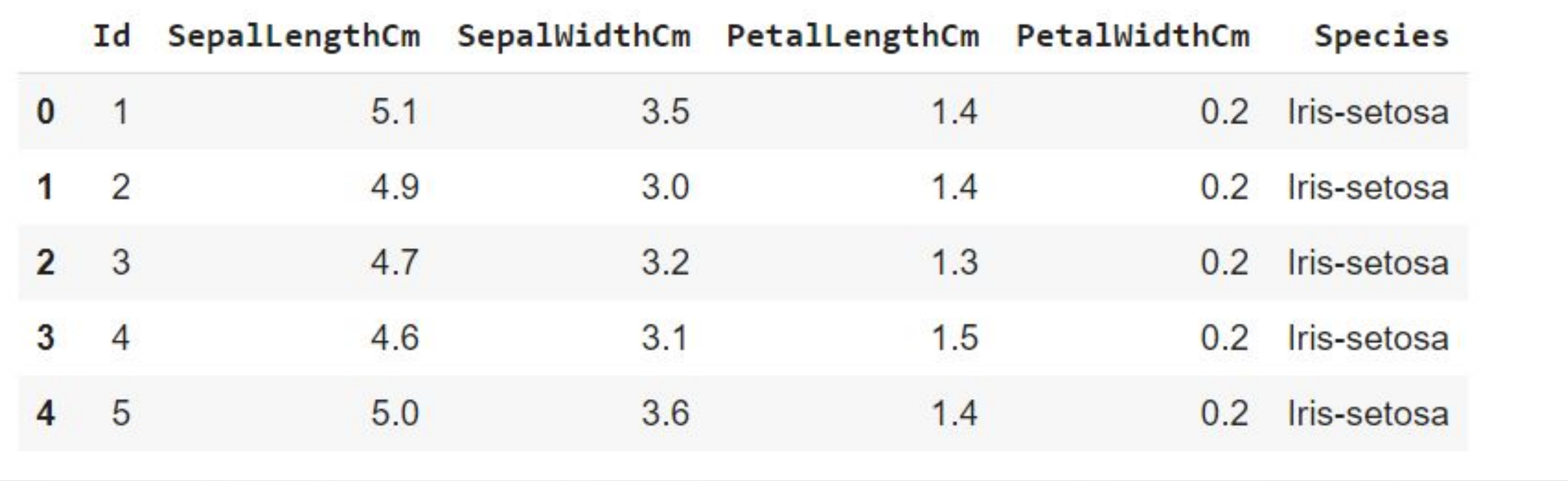
Name = GAUTAM Kumar
roll =B19EE031

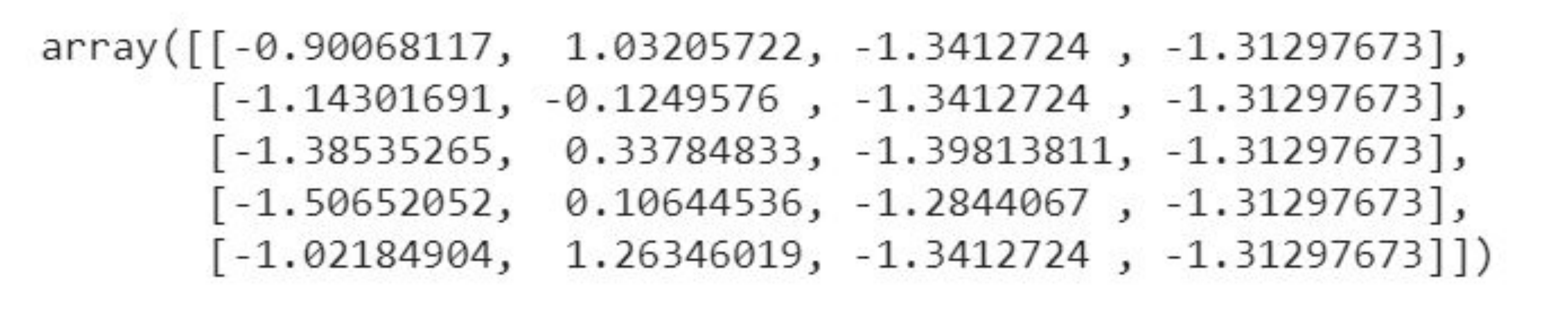
Lab 8 report

Head of dataset



	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

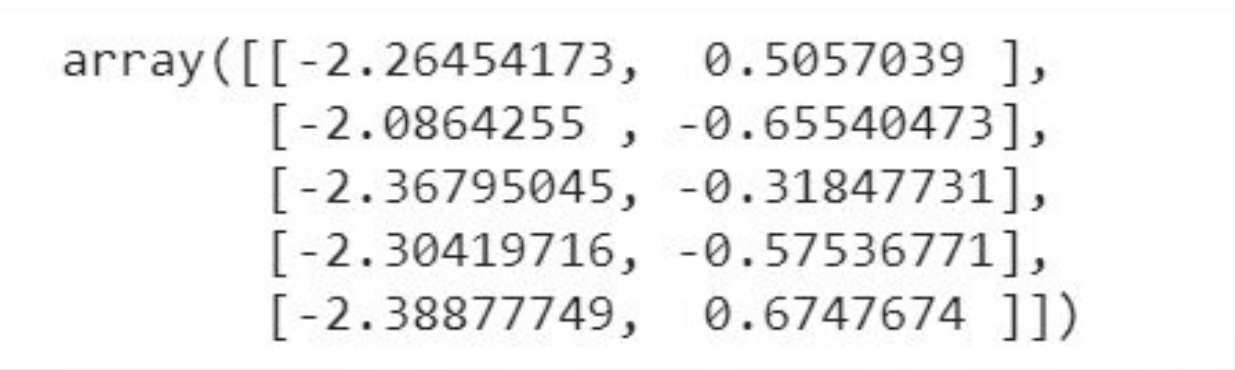
Dataset after applying standard scaler



```
array([[-0.90068117,  1.03205722, -1.3412724 , -1.31297673],
       [-1.14301691, -0.1249576 , -1.3412724 , -1.31297673],
       [-1.38535265,  0.33784833, -1.39813811, -1.31297673],
       [-1.50652052,  0.10644536, -1.2844067 , -1.31297673],
       [-1.02184904,  1.26346019, -1.3412724 , -1.31297673]])
```

ANALYSIS

PCA feature reduction by 90%



```
array([[-2.26454173,  0.5057039 ],
       [-2.0864255 , -0.65540473],
       [-2.36795045, -0.31847731],
       [-2.30419716, -0.57536771],
       [-2.38877749,  0.6747674 ]])
```

- ***Model :*** `PCA(copy=True, iterated_power='auto', n_components=0.9, random_state=None, svd_solver='auto', tol=0.0, whiten=False)`
- ***Ratio of variance in each column :*** `array([0.72770452, 0.23030523])`
- ***Contribution of each feature in PCA1 and PCA2 respectively:*** `array([[ 0.52237162, -0.26335492, 0.58125401, 0.56561105], [ 0.37231836, 0.92555649, 0.02109478, 0.06541577]])`
- ***Variance of both columns:*** `array([2.93035378, 0.92740362])`

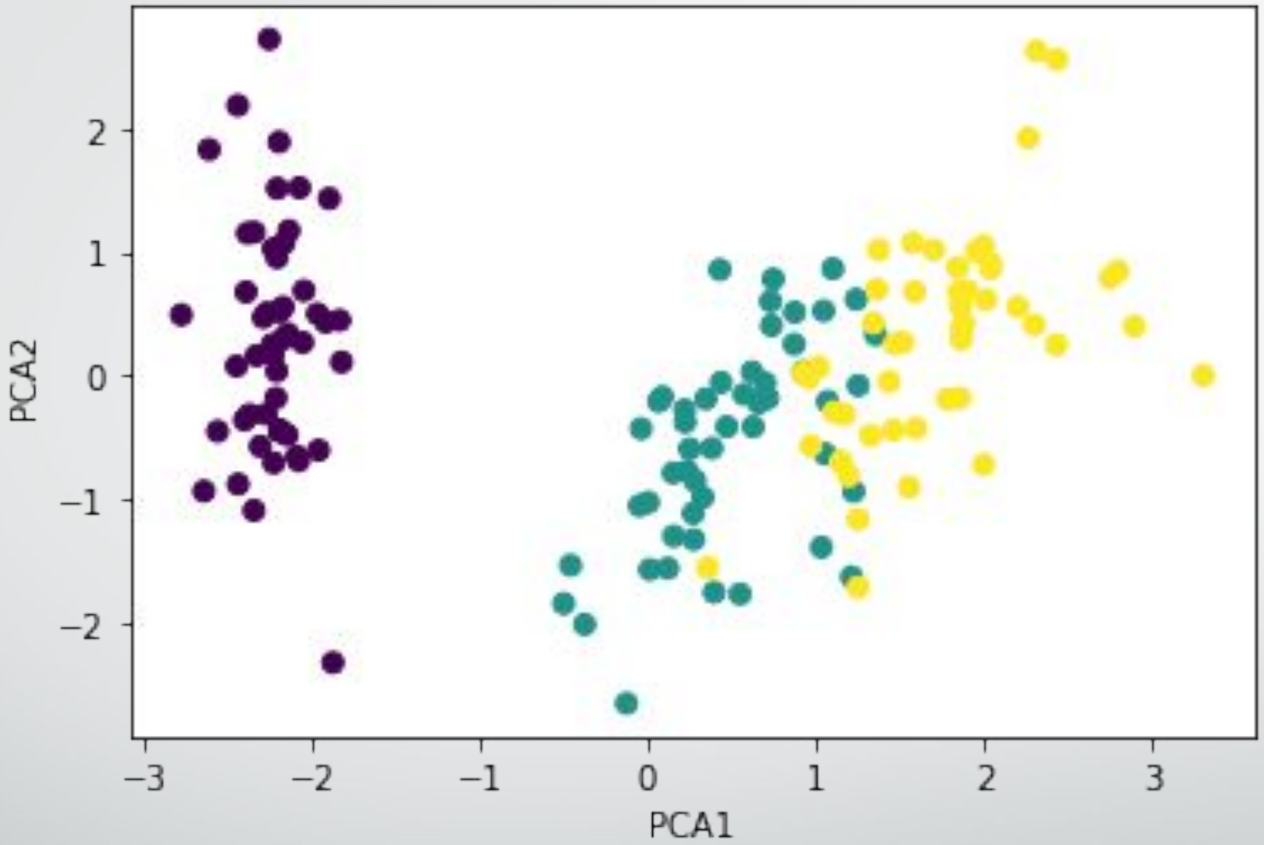
ANALYSIS

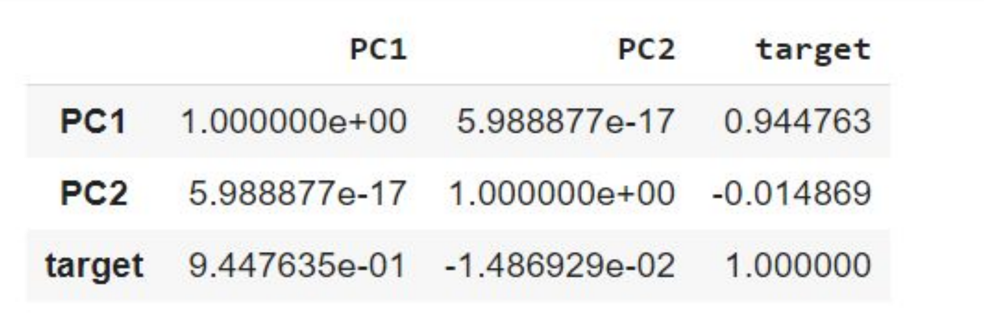
- ***Eigen vector of each principal components PC1 and PC2:*** `array([`

  ```
      [ 0.52237162, -0.26335492, 0.58125401, 0.56561105],
      [ 0.37231836, 0.92555649, 0.02109478, 0.06541577]])
  ```

- ***So we can see that feature 1,3 and 4 have higher values of eigenvalues for PC1 and the eigenvalues are positive. Feature 2 has low negative eigenvalue.***
- ***For second principal component, feature2 has very high contribution.***
- ***We can also see that both PC1 and PC2 are orthogonal to each other.***

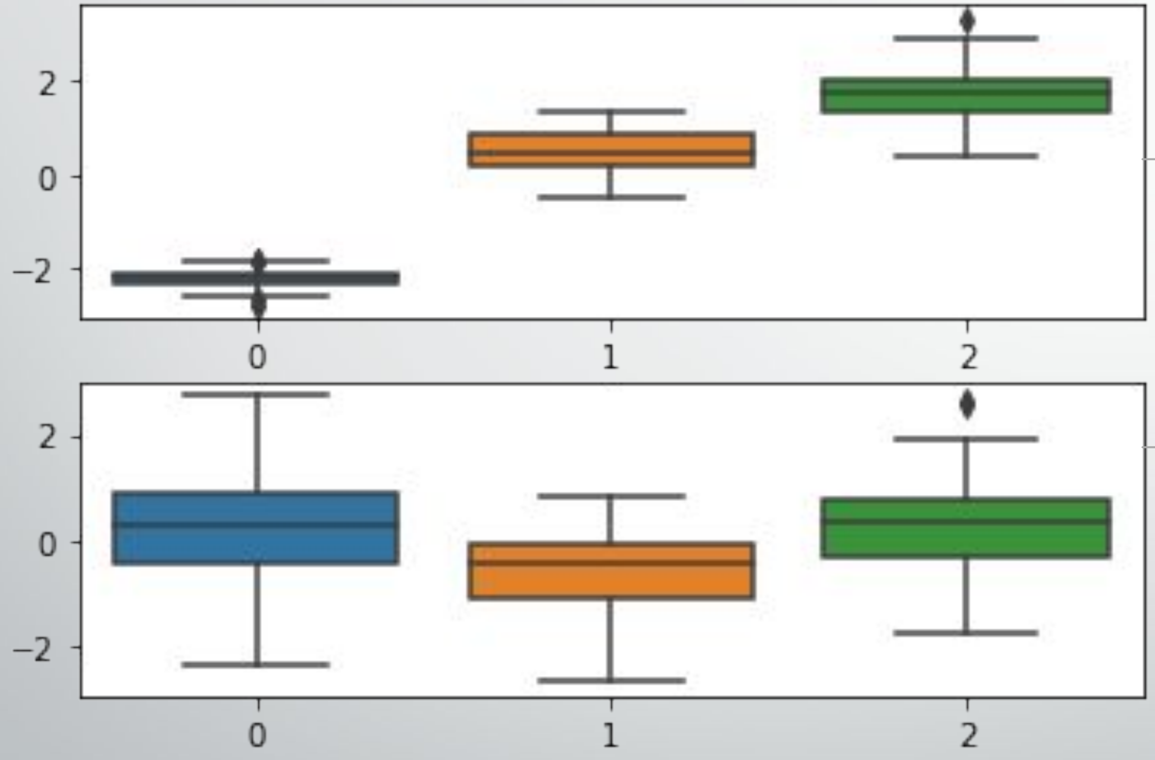
CLUSTERING BY PCA





	PC1	PC2	target
PC1	1.000000e+00	5.988877e-17	0.944763
PC2	5.988877e-17	1.000000e+00	-0.014869
target	9.447635e-01	-1.486929e-02	1.000000

Analysis of principal components of PCA



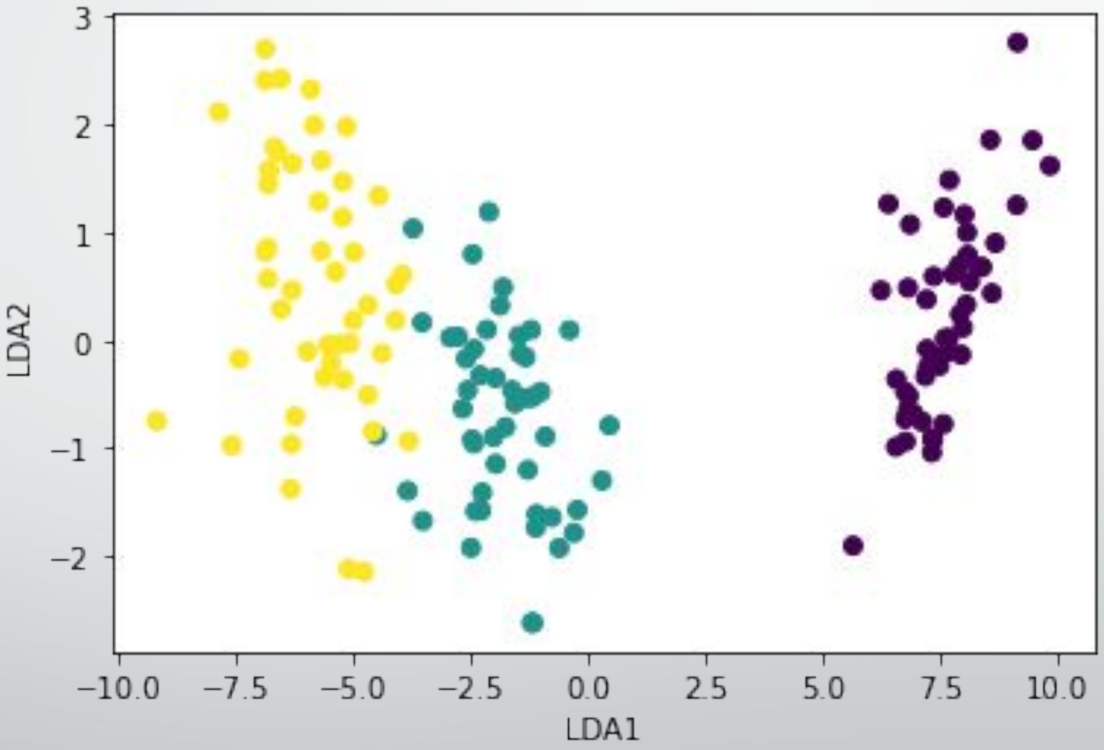
***CLASS**.*

We can see the variation corresponding to each class in the second component of resultant PCA is higher than the first component.

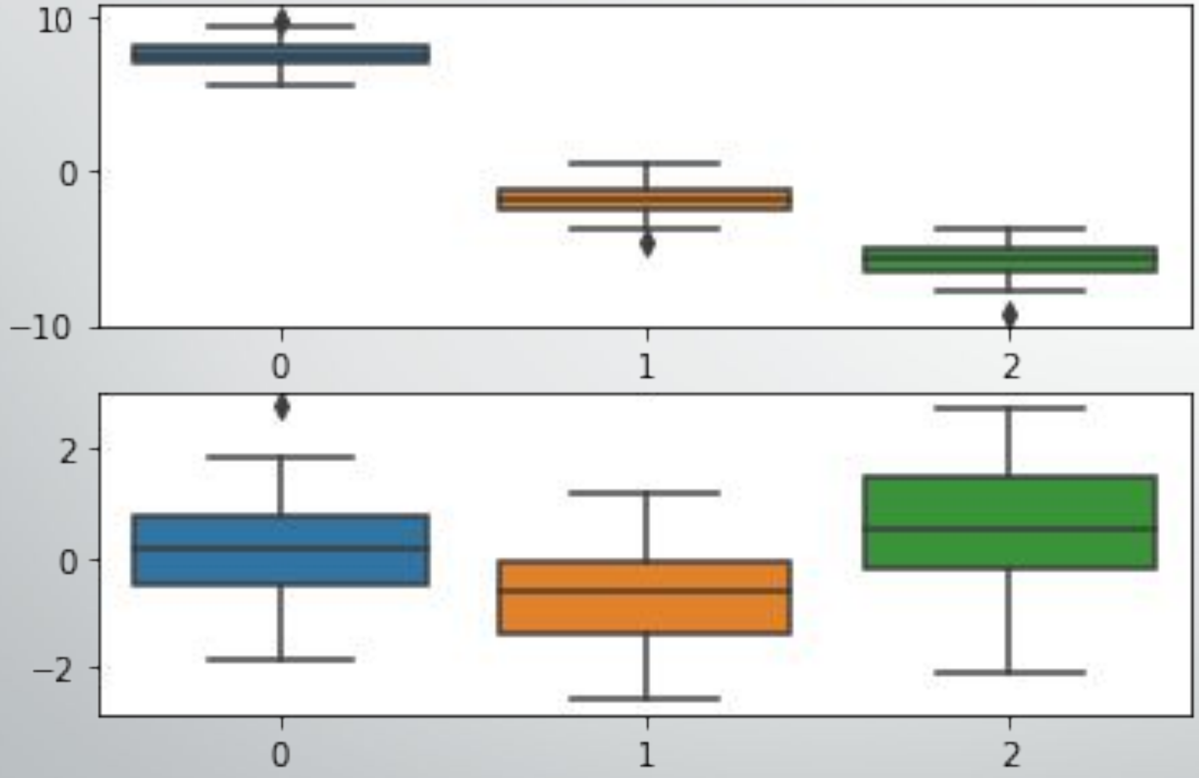
As class is increasing from 0 to 2, the average value of PCA1 is also increasing (high correlation).

PCA2 is neither increasing nor decreasing with increase in class value. Hence, we can say that PCA1 is the most important principal component for clustering the dataset.

CLUSTERING BY LDA



Analysis of principal components of LDA



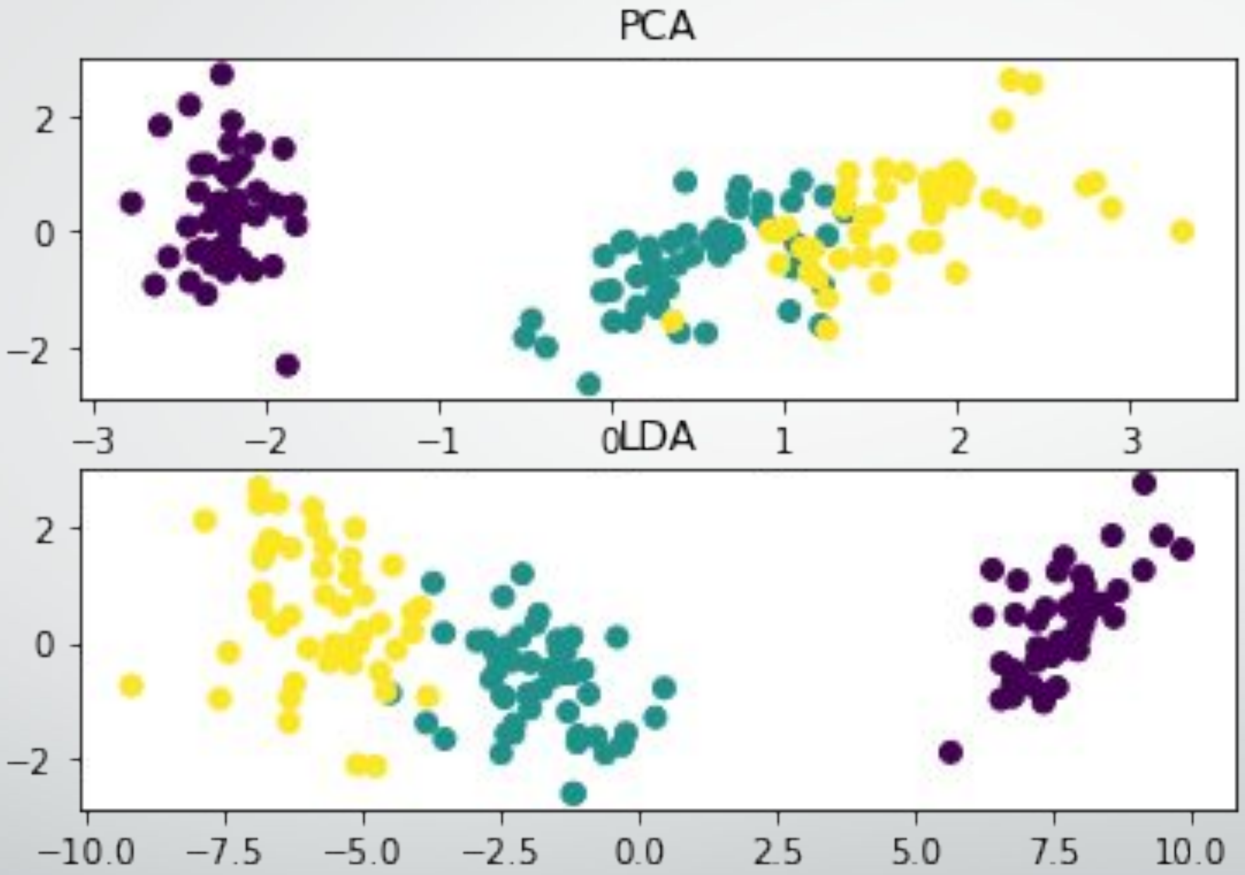
CLASS.

We can see the variation in the second component of resultant LDA is lower than the first component.

As class is increasing from 0 to 2, the average value of LDA1 is decreasing.

LDA2 is neither increasing nor decreasing with increase in class value. Hence, we can say that LDA1 is the most important component for clustering the dataset.

PCA vs LDA



Accuracy of PCA = 0.8
Accuracy of LDA = 0.9

Analysis

- *The LDA model preformed better than PCA model.*
- *LDA takes the output into account while PCA perform feature reduction only based on input variables.*
- *In PCA 90% data conservation came through 2 eigen vectors.*
- *In nutshell we can say that PCA is concerned with differences in x-values while LDA is concerned with differences in x-values with respect to y.*
- *Both the models are used for feature reduction.*

Feature Selection

- *Method 1:* *RFE (Recursive Feature Ellimination)*
- *Parameters:*

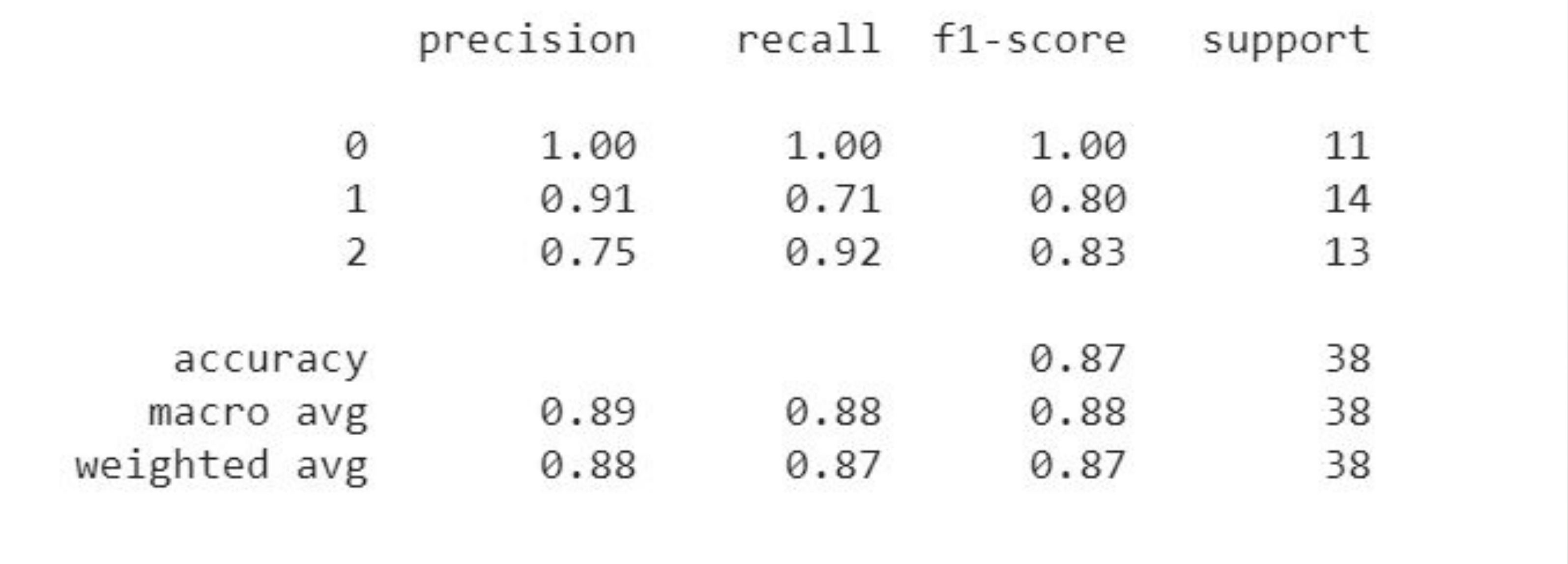
```
RFE(estimator=LinearDiscriminantAnalysis(n_components=2,
priors=None, shrinkage=None, solver='svd',
store_covariance=False, tol=0.0001), n_features_to_select=2,
step=1, verbose=0)
```

- *Selected Features :* `['SepalWidthCm', 'PetalWidthCm’]`

Feature Selection Method 2

- *Model 2: Select k-best, score function: Chi Squared*
- *Parameters :* `SelectKBest(k=2, score_func=<function chi2)`
- *Score of each feature :* `array([ 10.81782088, 3.59449902, 116.16984746, 67.24482759])`
- *Selected Features :* `['PetalLengthCm', 'PetalWidthCm']`

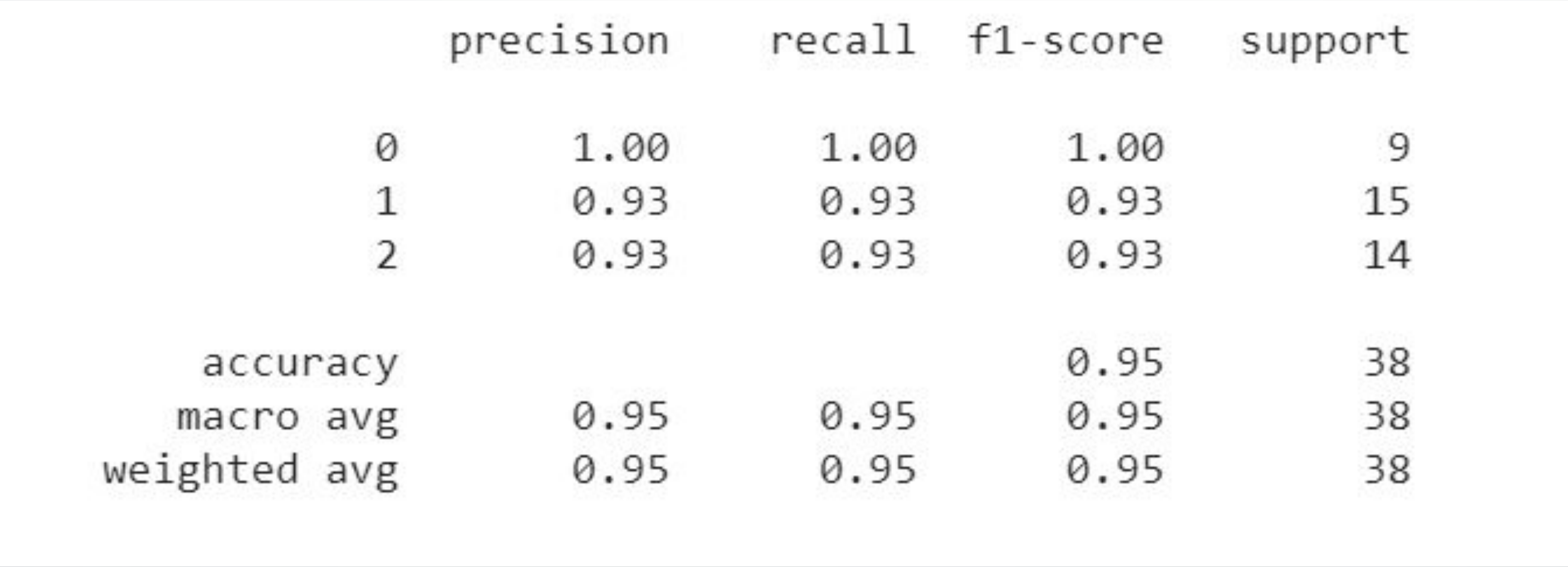
Classification report of feature selection by RFE model



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        11
           1       0.91      0.71      0.80        14
           2       0.75      0.92      0.83        13

    accuracy                           0.87        38
   macro avg       0.89      0.88      0.88        38
weighted avg       0.88      0.87      0.87        38
```

Accuracy : 0.9210526315789473, **f1-score**: 0.87

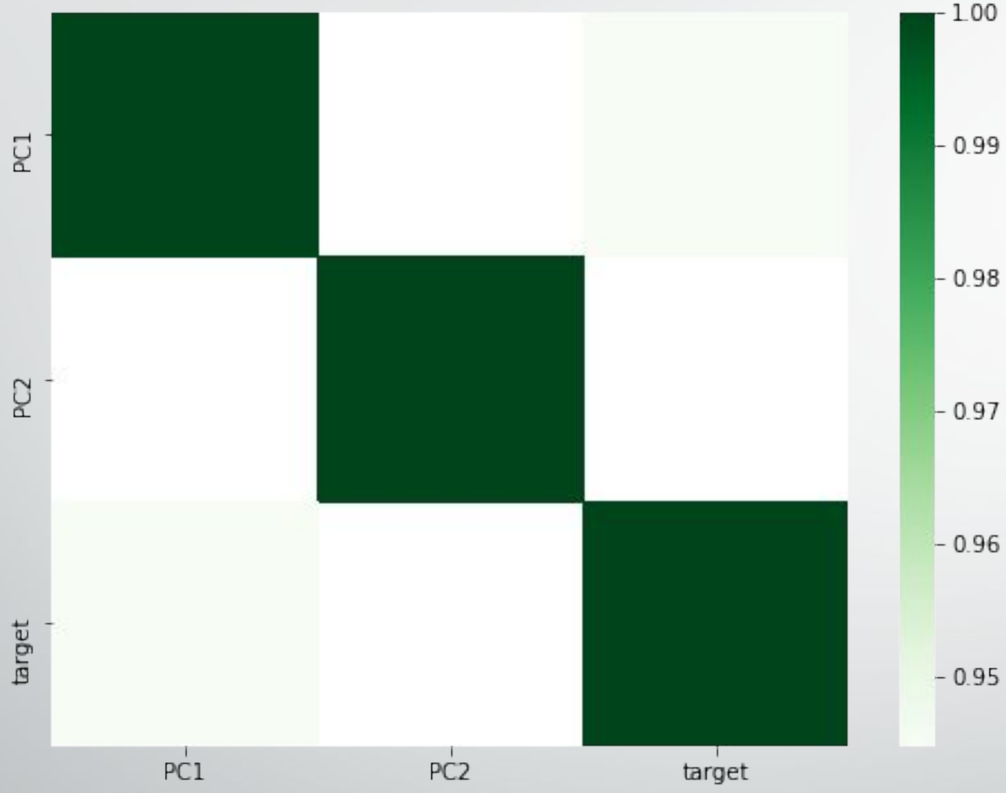
Classification report of feature selection by Select-KBest model

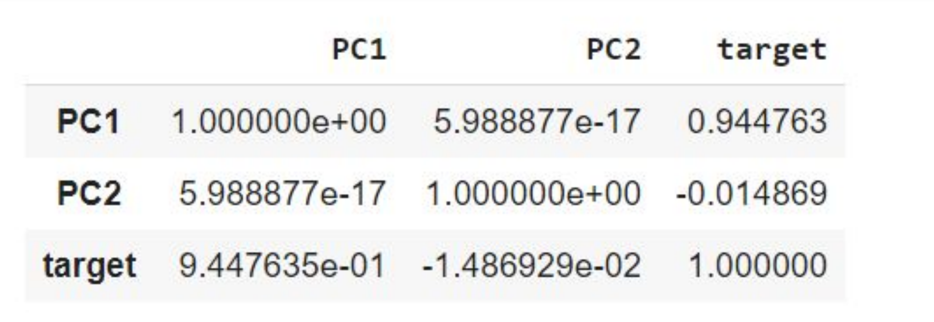


```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         9
           1       0.93      0.93      0.93        15
           2       0.93      0.93      0.93        14

    accuracy                           0.95        38
   macro avg       0.95      0.95      0.95        38
weighted avg       0.95      0.95      0.95        38
```

Accuracy : 0.9736842105263158, **f1-score:** 0.95

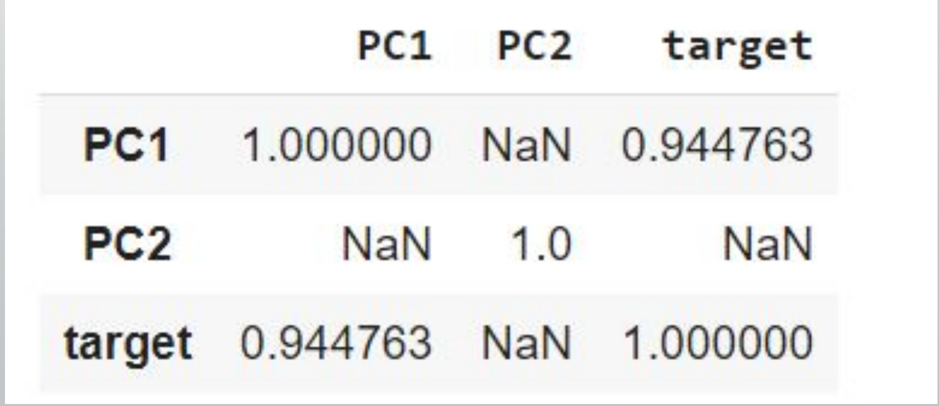
For PCA model correlation greater than 0.7





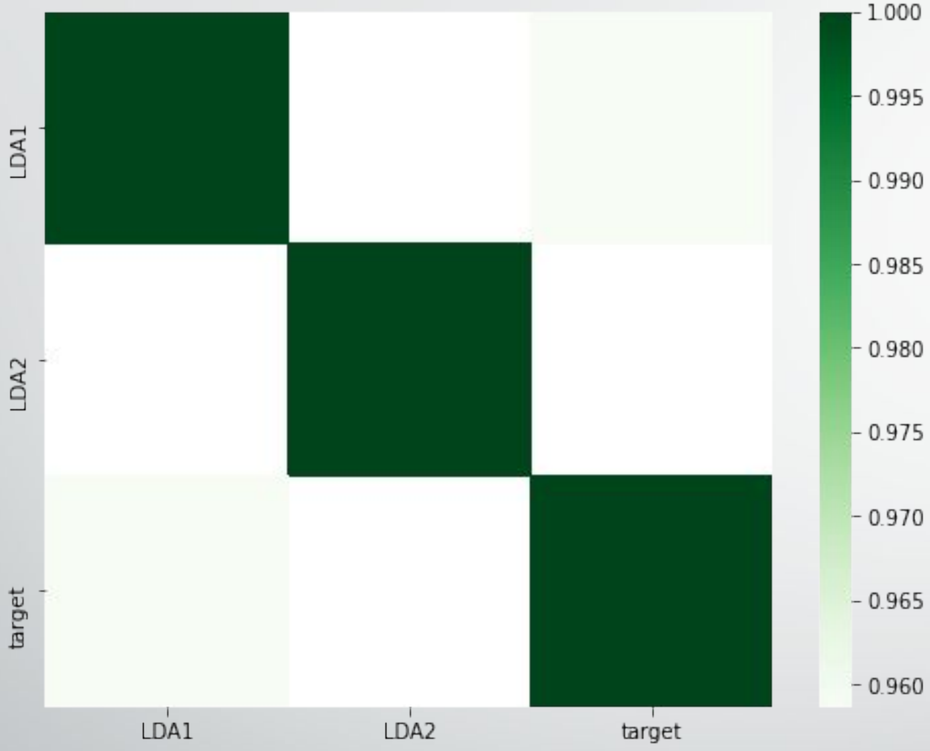
	PC1	PC2	target
PC1	1.000000e+00	5.988877e-17	0.944763
PC2	5.988877e-17	1.000000e+00	-0.014869
target	9.447635e-01	-1.486929e-02	1.000000

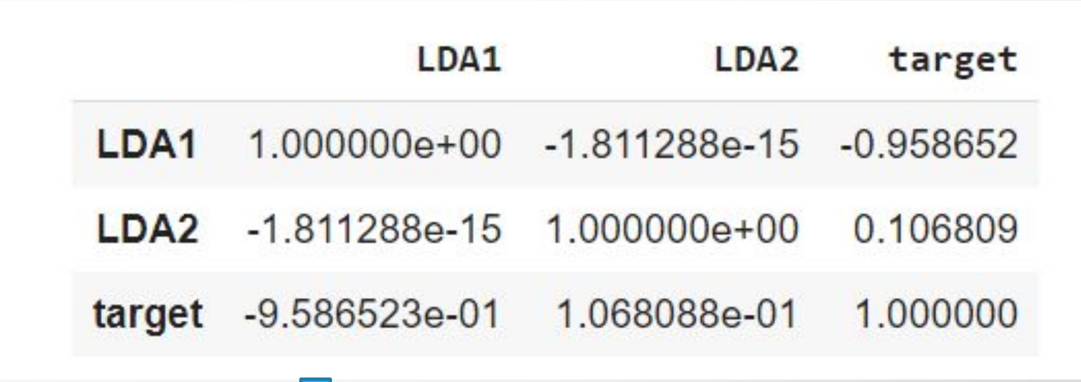
Applying 70% filter



	PC1	PC2	target
PC1	1.000000	NaN	0.944763
PC2	NaN	1.0	NaN
target	0.944763	NaN	1.000000

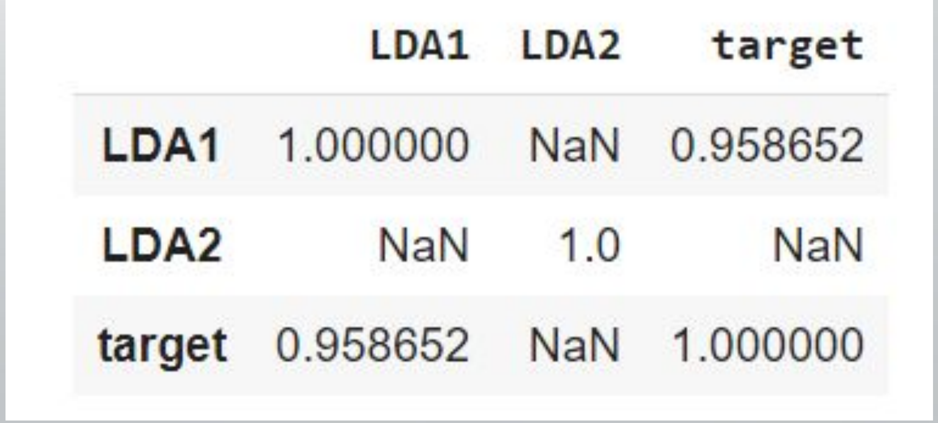
For LDA model correlation greater than 0.7





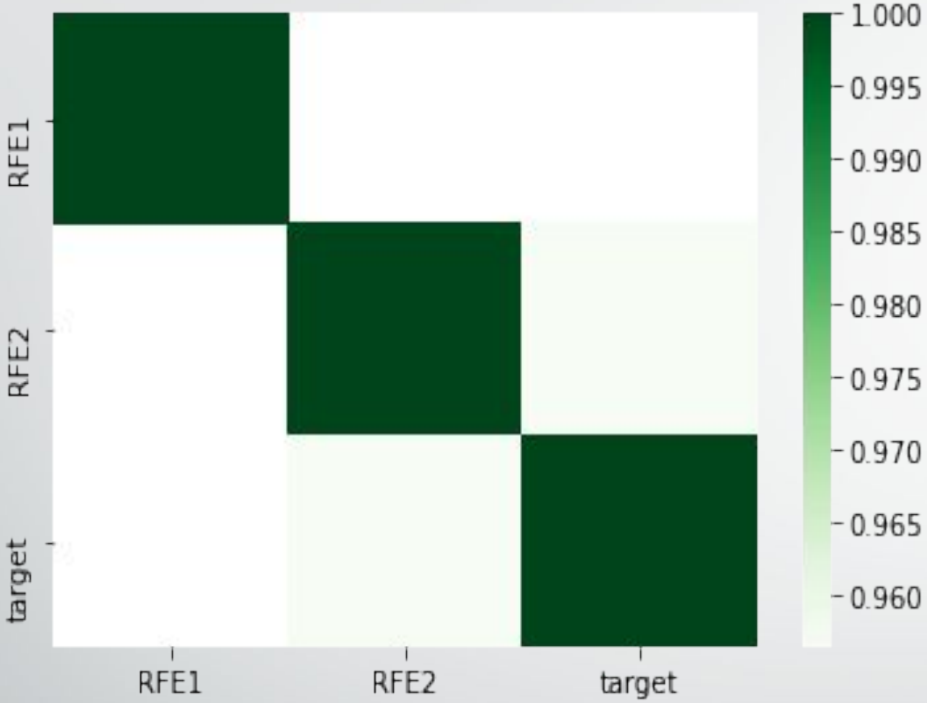
	LDA1	LDA2	target
LDA1	1.000000e+00	-1.811288e-15	-0.958652
LDA2	-1.811288e-15	1.000000e+00	0.106809
target	-9.586523e-01	1.068088e-01	1.000000

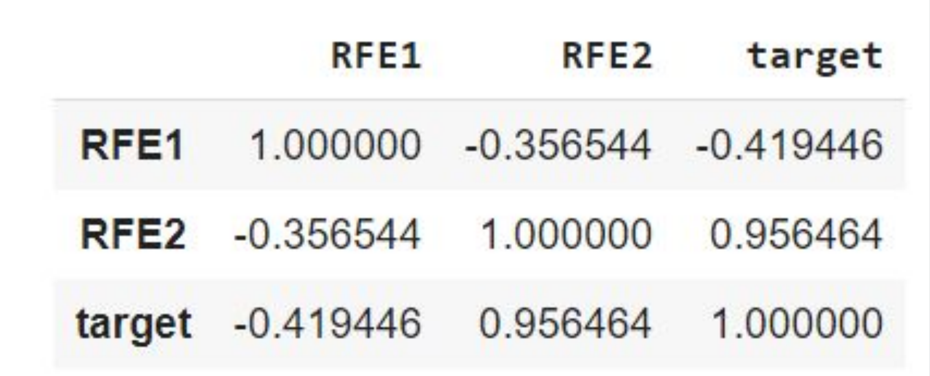
Applying 70% filter



	LDA1	LDA2	target
LDA1	1.000000	NaN	0.958652
LDA2	NaN	1.0	NaN
target	0.958652	NaN	1.000000

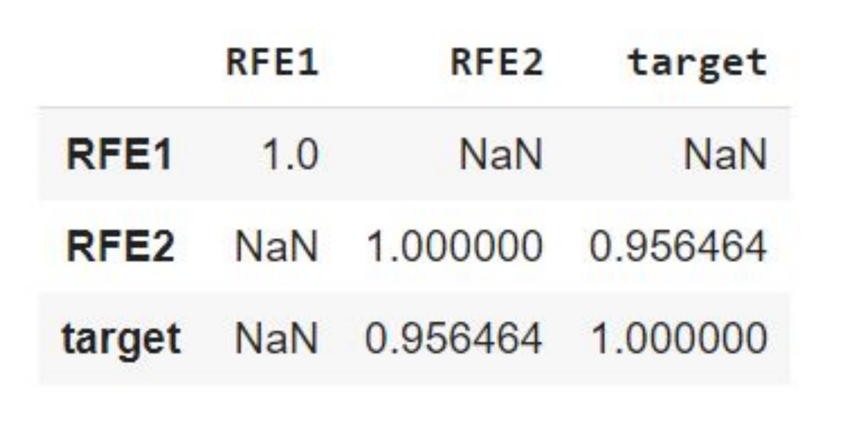
For RFE model correlation greater than 0.7





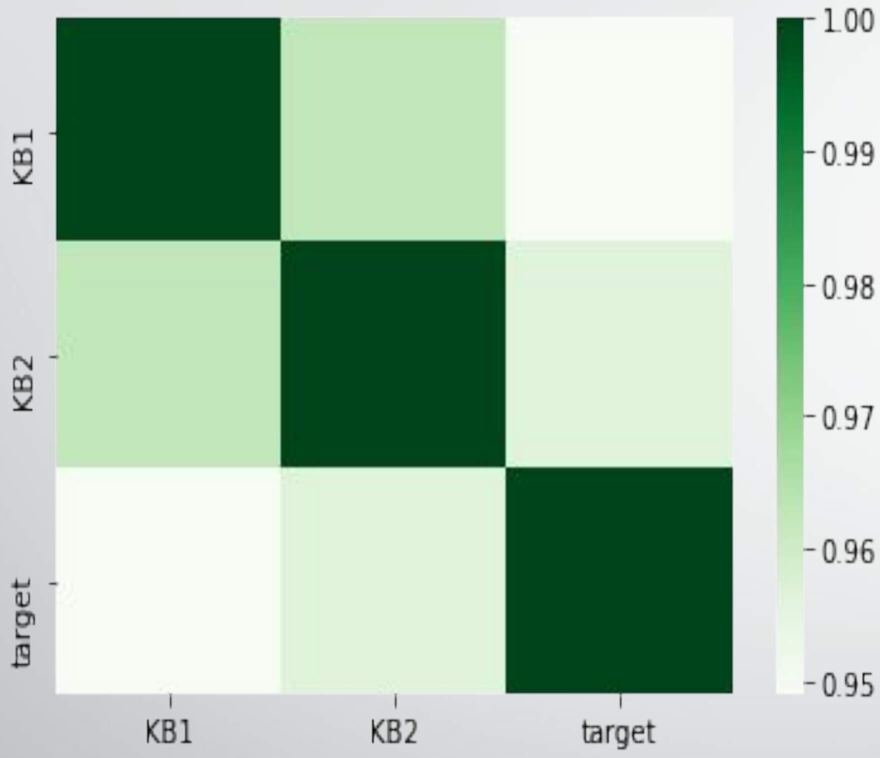
	RFE1	RFE2	target
RFE1	1.000000	-0.356544	-0.419446
RFE2	-0.356544	1.000000	0.956464
target	-0.419446	0.956464	1.000000

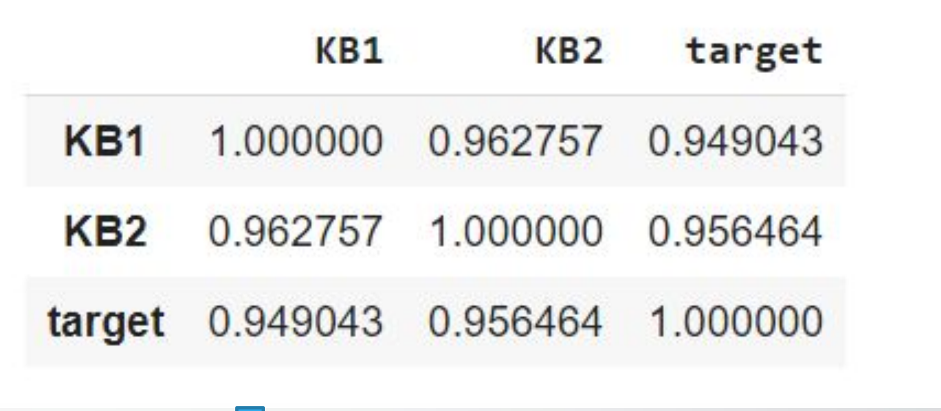
Applying 70% filter



	RFE1	RFE2	target
RFE1	1.0	NaN	NaN
RFE2	NaN	1.000000	0.956464
target	NaN	0.956464	1.000000

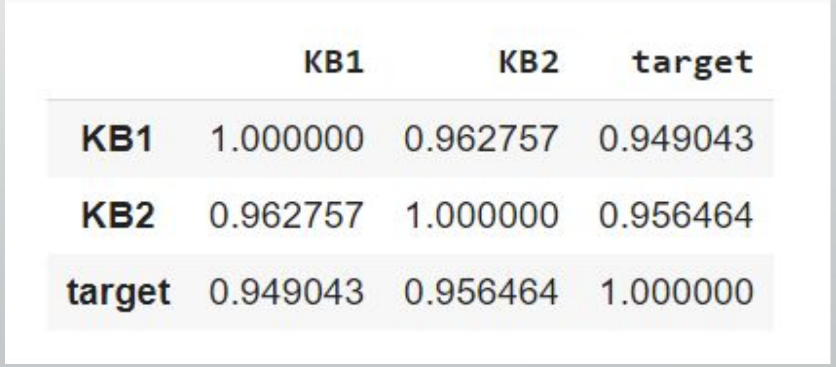
For kbest model correlation greater than 0.7





	KB1	KB2	target
KB1	1.000000	0.962757	0.949043
KB2	0.962757	1.000000	0.956464
target	0.949043	0.956464	1.000000

Applying 70% filter



	KB1	KB2	target
KB1	1.000000	0.962757	0.949043
KB2	0.962757	1.000000	0.956464
target	0.949043	0.956464	1.000000

Unchanged

ANALYSIS

- *In PCA model, PC1 is more correlated to target than PC2. Hence, we can say that most important component of a PCA model is the first set of eigen vectors.*
- *Similarly, in LDA model, first axis is more correlated to target than second. Hence, the priority decreases with increasing column number.*
- *In RFE second attribute had higher correlation but this cannot be generalised for such models.*
- *In k-best model, both the selected features had higher correlation than 70%.*
- *K-best model performed better than RFE in terms of both parameters, accuracy as well as f1-score.*

CONCLUSION

- *We learnt about dimension reduction.*
- *We analyzed PCA (Principal Component Analysis) model.*
- *We compared PCA with LDA model and found their differences and similarities.*
- *We learnt about feature selection.*
- *We analysed and compared RFE model and k-best model.*
- *Lastly, we analysed the correlation matrix of all the models and their principal components.*