Name: Prathamesh Uravane

UID: 122016187
Project Title: Feature Selection and Dimensionality Reduction

1. Objective

To Understand explore directionality reduction techniques in for classification problems with feature selection .

Involved datasets:

1. **Pollution_dataset**

2. Dataset Description

Heart Disease Dataset

- **Records:** 5000 samples
- **Attributes:** 9 predictive variables
- **Target:** 4 classes

3. SVM and Gaussian Naïve Bayes

Two models were trained on the Heart Disease dataset:

- **SVM**
- **GaussinaNB**

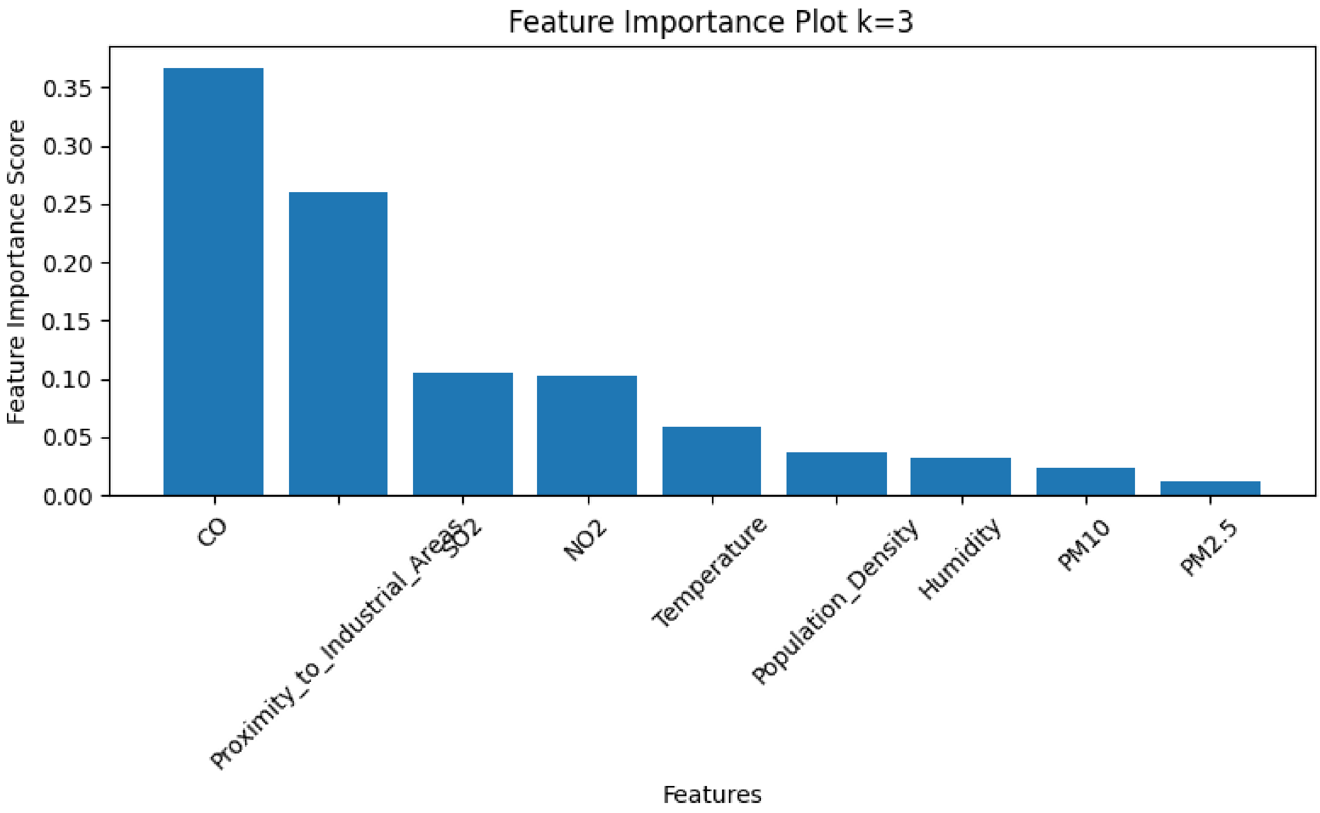
Results Discussion

1. Univariate feature selection method

1. Used method – Chi square
2. Selected k features where k = {1,2,3}
3. Selected feature names are as follows:
   - 3.1. k=1 → ['CO']
   - 3.2. k=2 → ['CO', 'Proximity_to_Industrial_Areas']
   - 3.3. k=3 → ['CO', 'Proximity_to_Industrial_Areas', 'SO2']

2. Feature importance score to select features

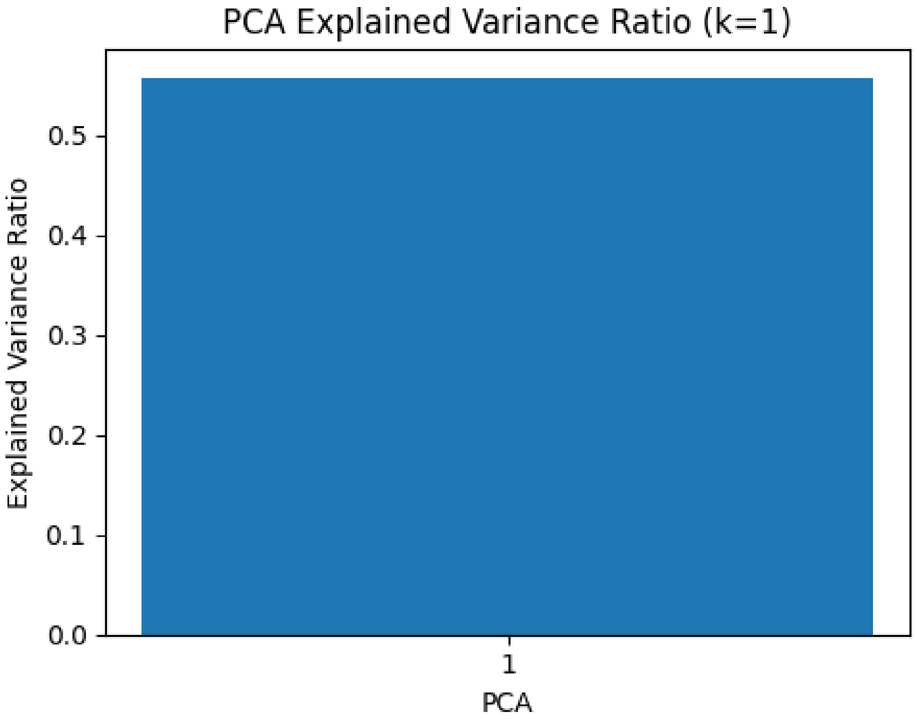
1. Classifier random forest (100 trees)
2. Selected k features where k = {1,2,3}
3. Selected feature names are as follows:
   3.1. k=1 → ['CO']
   3.2. k=2 → ['CO', 'Proximity_to_Industrial_Areas']
   3.3. k=3 → ['CO', 'Proximity_to_Industrial_Areas', 'SO2']

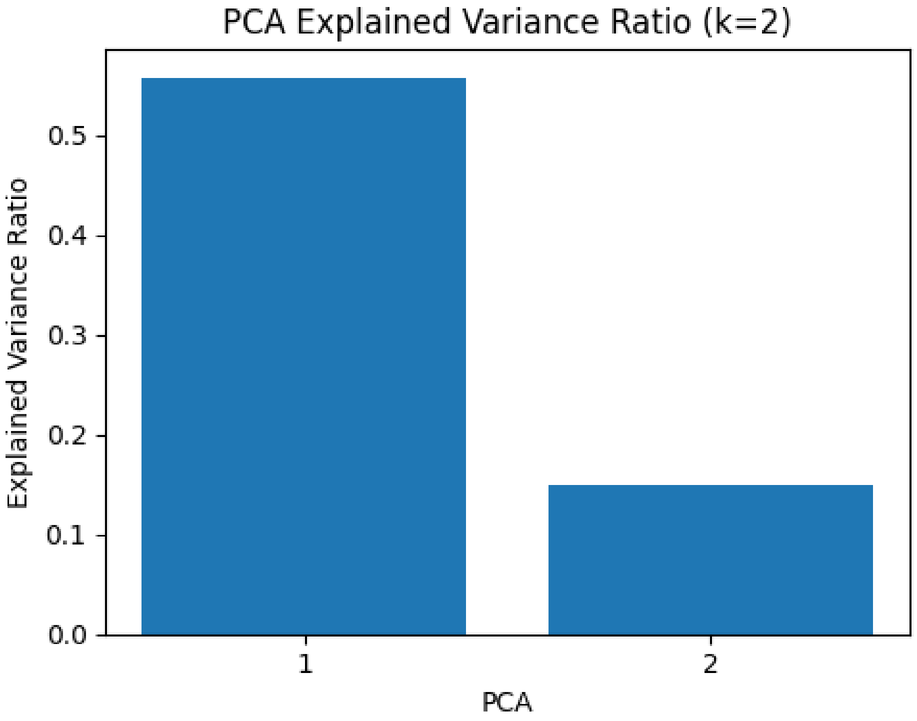


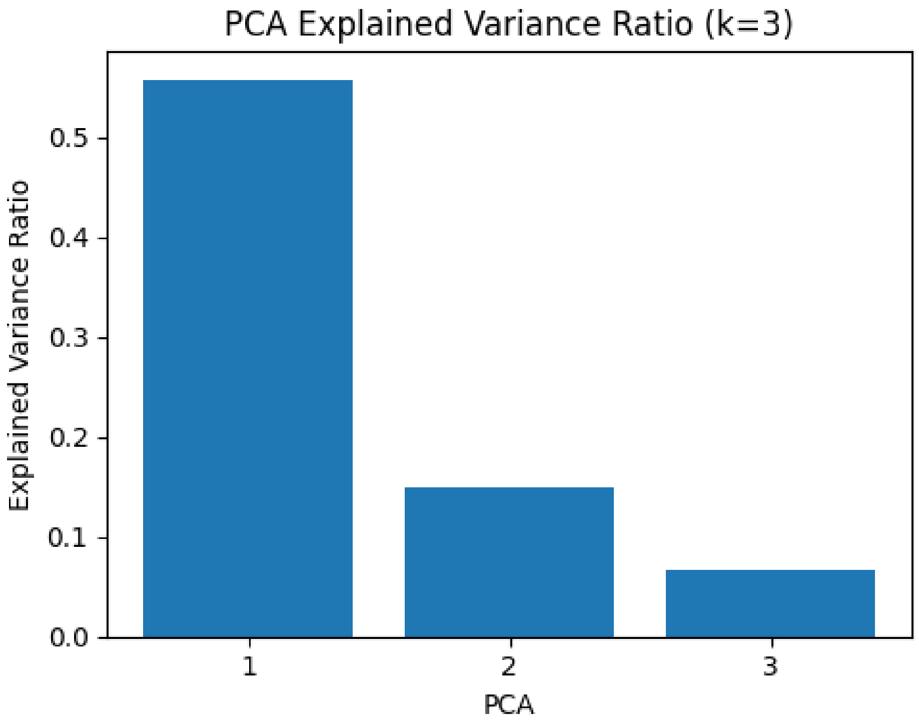
3. Principle component analysis(Un-supervised)

1. Before transforming into PCA standard scalling has been applied
2. After PCA trasnformation experince vaiance ration has been calculated
3. Experinence variance ratio is as follows for k no. of componants ‘k = {1,2,3}’

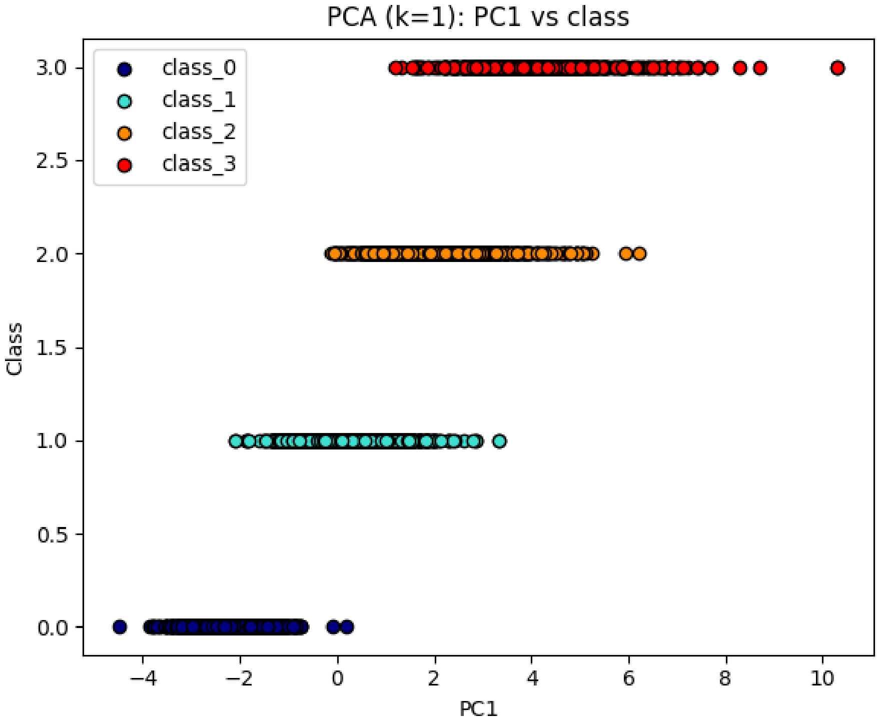
   1. PCA (k=1) explained variance ratio: [0.55741525]
   2. PCA (k=2) explained variance ratio: [0.55741525 0.14969902]
   3. PCA (k=3) explained variance ratio: [0.55741525 0.14969902 0.06602705]

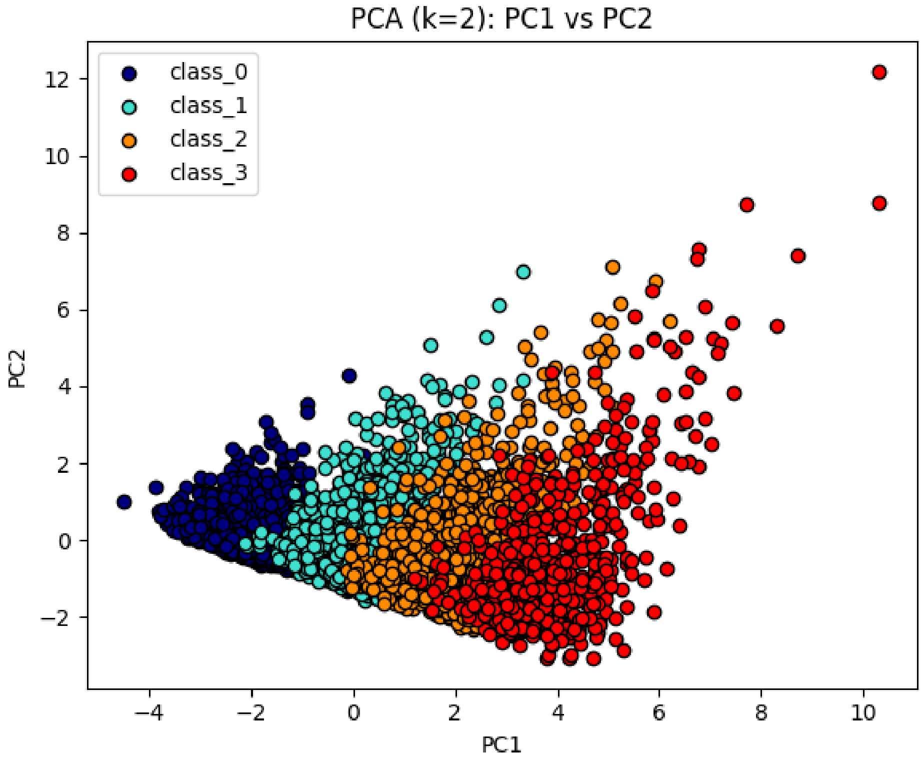

PCA Explained Variance Ratio (k=1)
Explained Variance Ratio
0.0
0.1
0.2
0.3
0.4
0.5
1
PCA

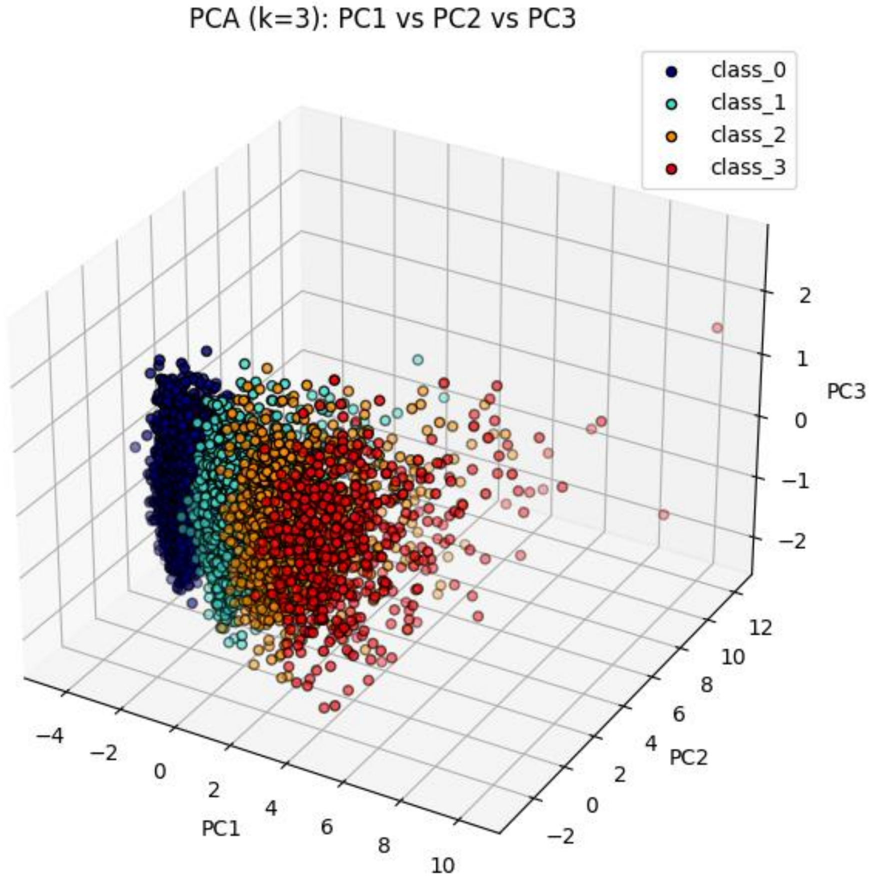

PCA Explained Variance Ratio (k=2)
Explained Variance Ratio
0.0
0.1
0.2
0.3
0.4
0.5
1
2
PCA



4. After PCA transformation data distribution looks like follows:



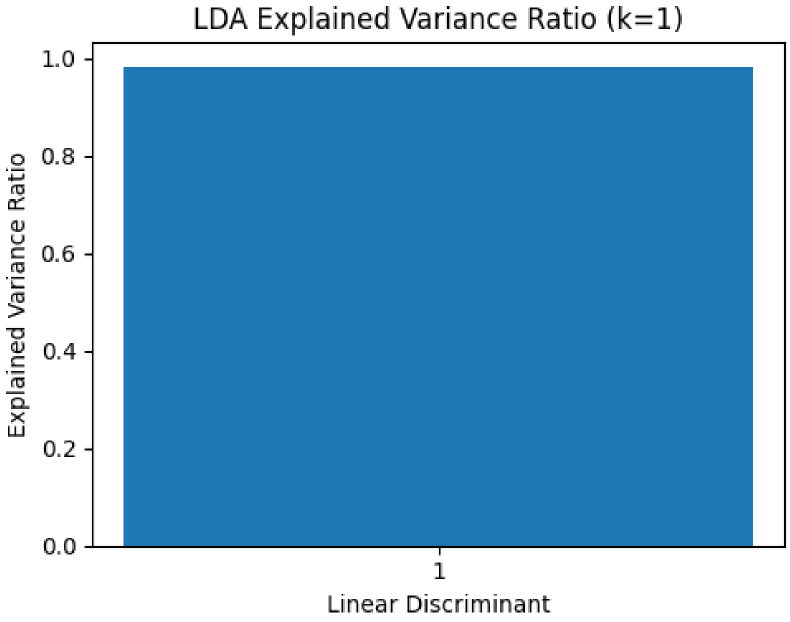

PCA (k=2): PC1 vs PC2

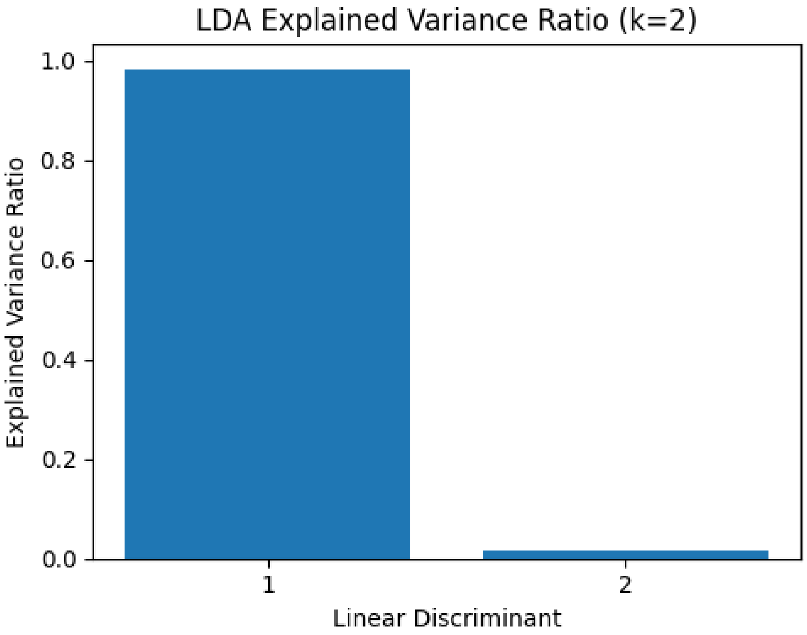

PCA (k=3): PC1 vs PC2 vs PC3

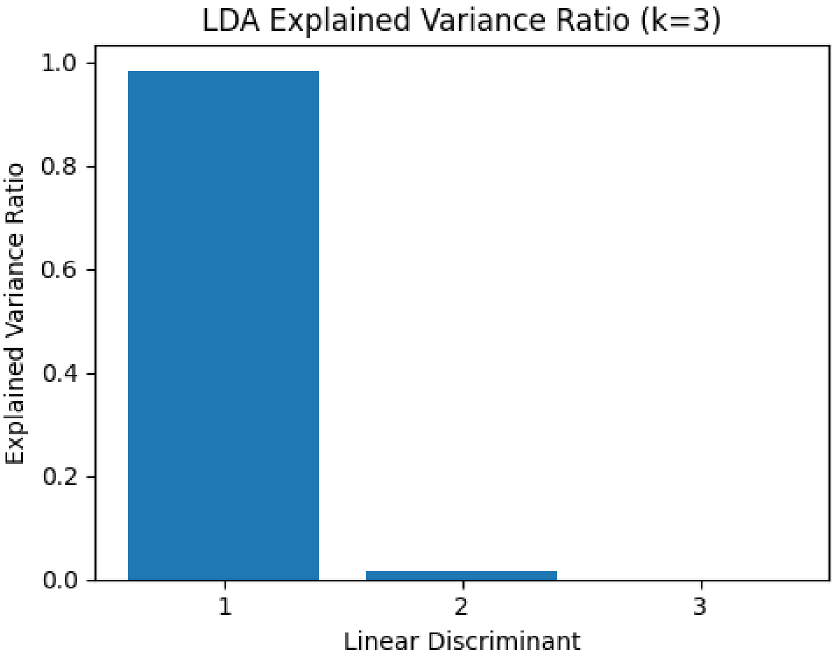
4. LDA (Linear Discriminant analysis) (Supervised)

1. Supervised learning method for feature transformation.
2. Before transforming into LDA standard scalling has been applied.
3. Target variable used to select features
4. 4. Experinence variance ratio is as follows for k no. of componants ‘k = {1,2,3}’

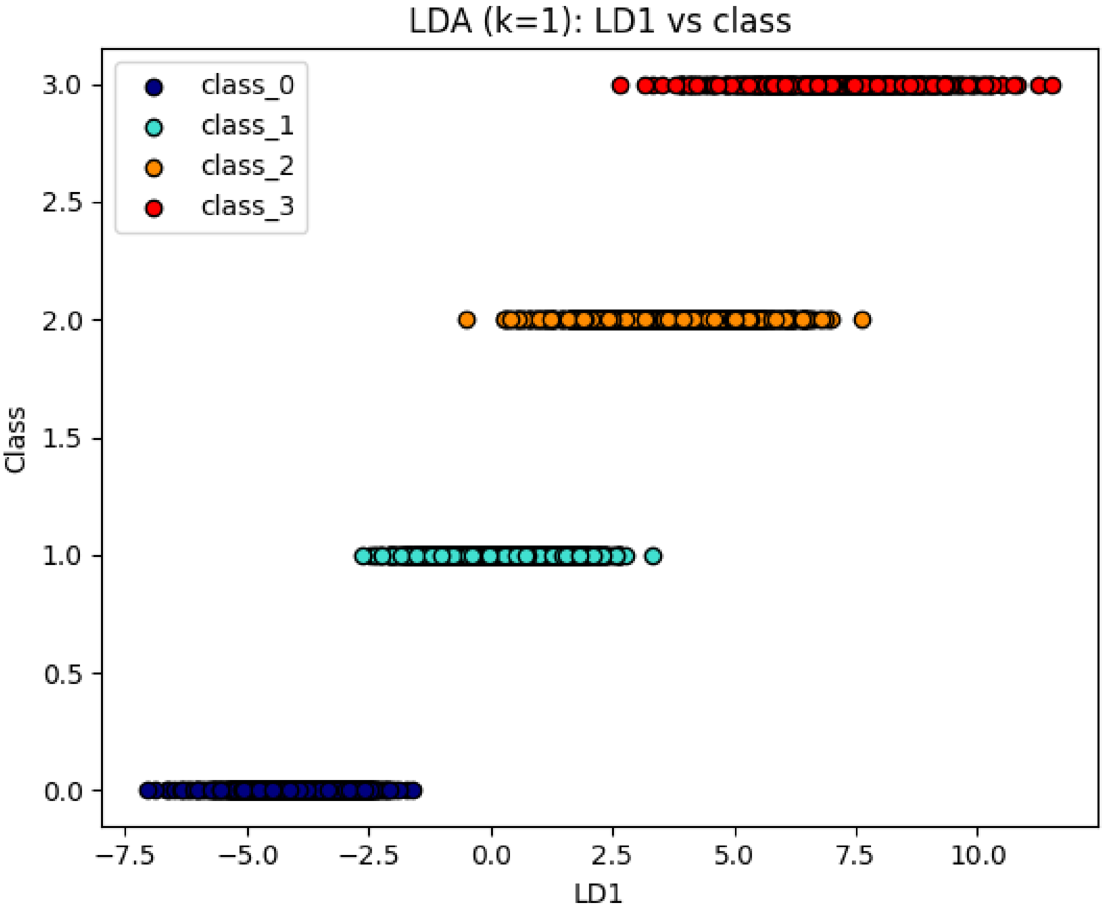
   1. LDA (k=1) explained variance ratio: [0.9822482]
   2. LDA (k=2) explained variance ratio: [0.9822482 0.01756347]
   3. LDA (k=3) explained variance ratio: [0.9822482 0.01756347 0.188336987]

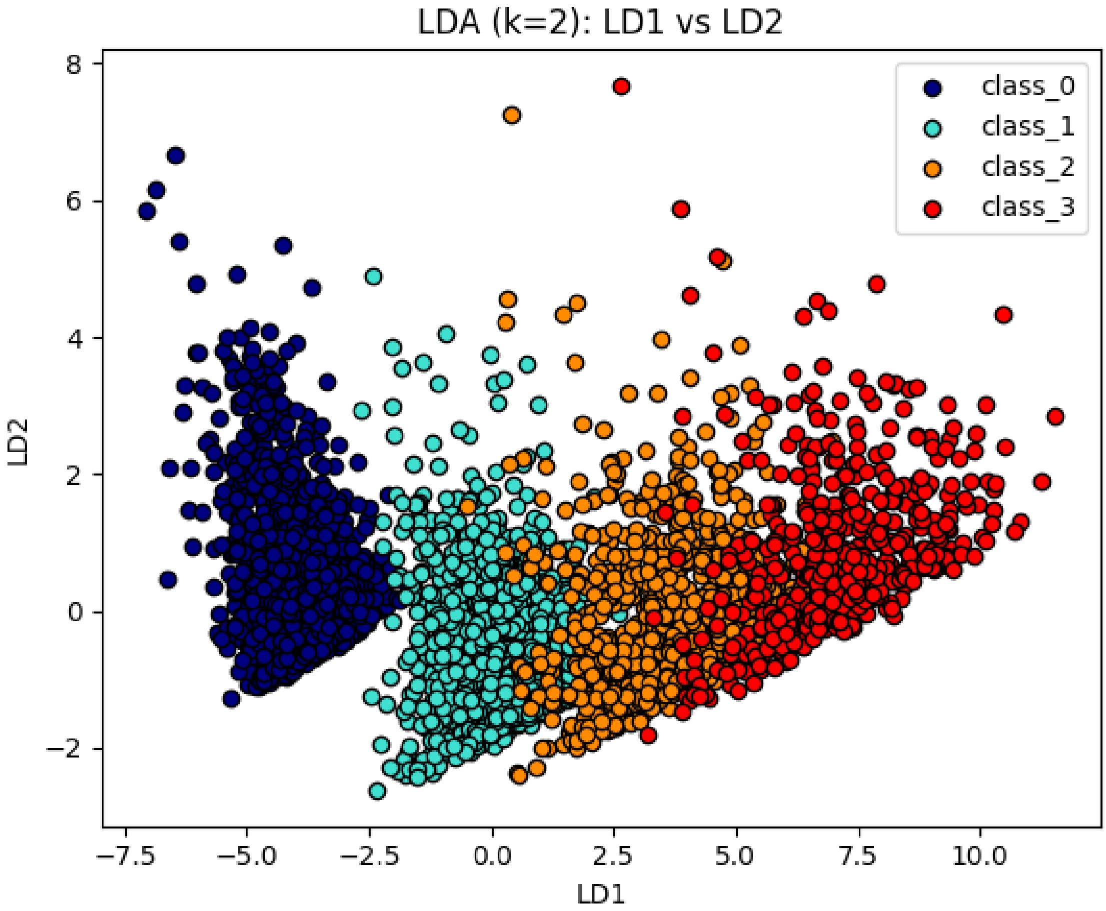


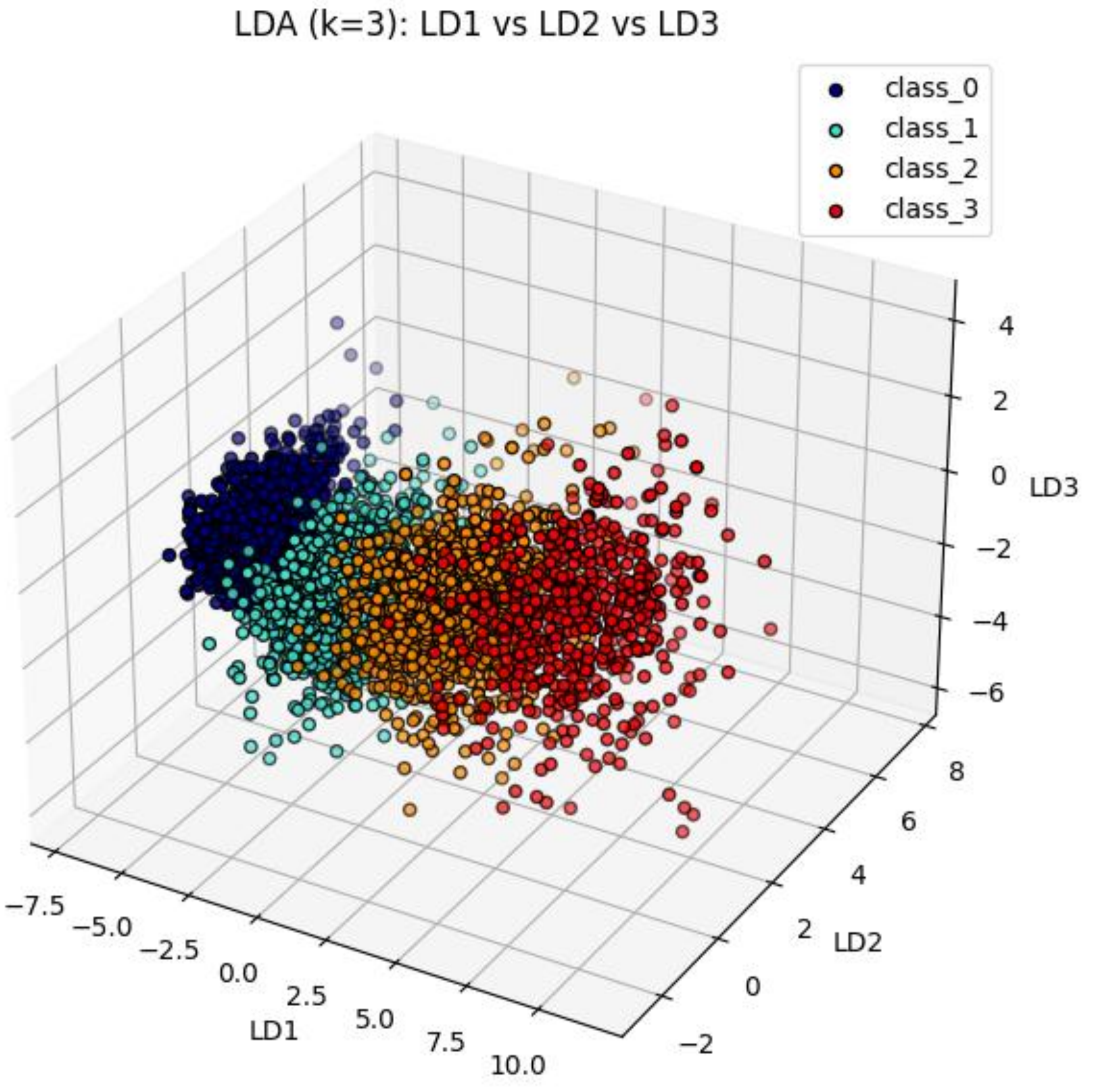




4. After PCA transformation data distribution looks like follows:




LDA (k=2): LD1 vs LD2
8
6
4
2
0
−2
LD2
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
LD1
class_0
class_1
class_2
class_3



Model training and evaluation

Two classifiers were considered:

- **Support Vector Machine (SVM)**
- **Gaussian Naive Bayes (GNB)**

50 randomly split training and testing datasets

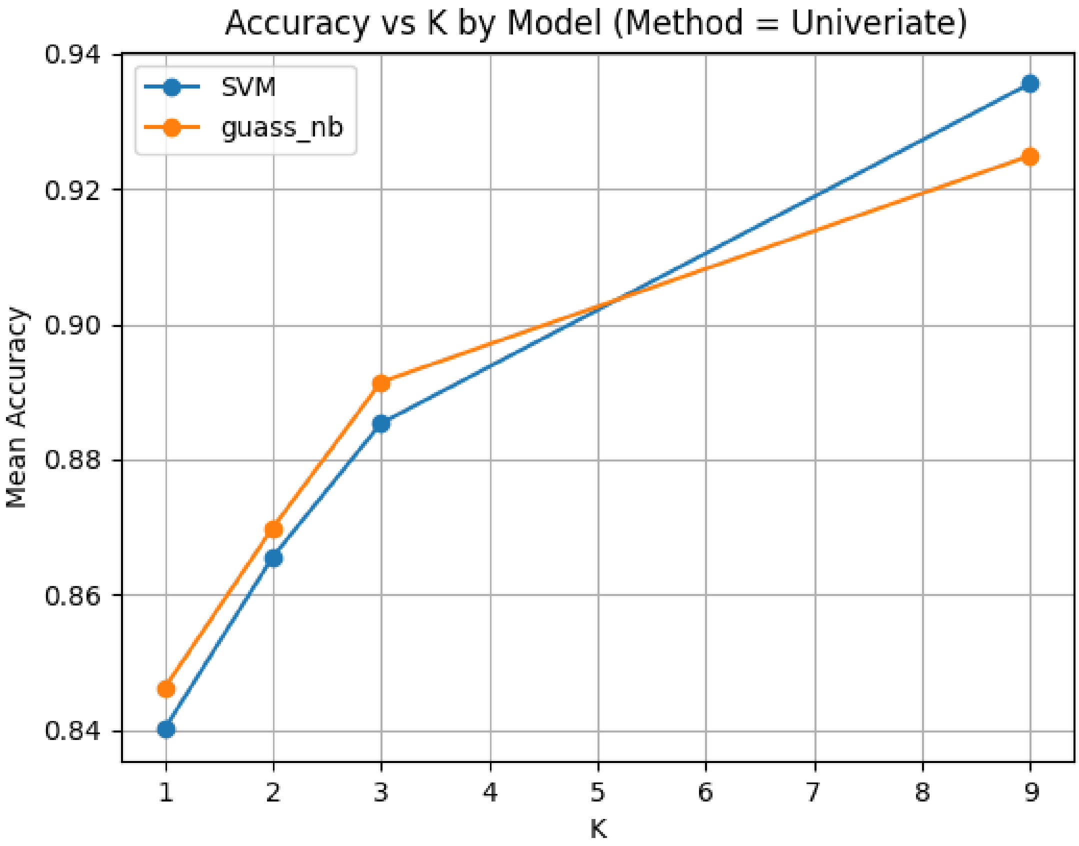
Overall Accuracy Patterns

Throughout all methods and models, the mean accuracy was **0.846**, with values ranging from **0.45** to **0.96**.

1. Univariate Selection

Average accuracy over both models increased strongly with K:

- **K=1:** 0. 843112
- **K=2:** 0. 867696
- **K=3:** 0. 888368
- **K=9:** 0. 930248 (All features)



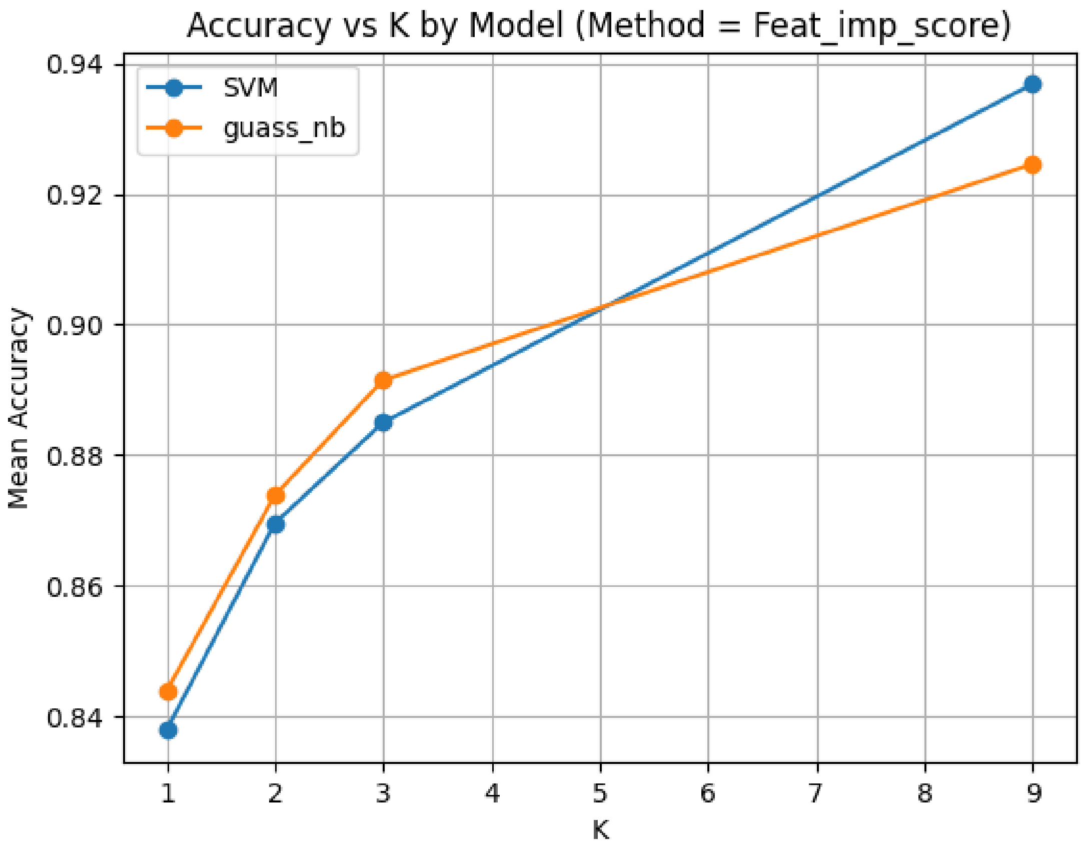
a. At small K (1–3) performance was lesser, while model performed better from additional features.

b. univariate ranking removes the interactions between variables. When only a few features are kept, there is possibility of losing important joint effects, which is a cause of relatively low accuracy.

2. Feature importance score

Average accuracy for “Feat_imp_score” was consistently high:

- **K=1:** 0. 840848
- **K=2:** 0. 871696
- **K=3:** 0. 888232
- **K=9:** 0. 930712 (All features)

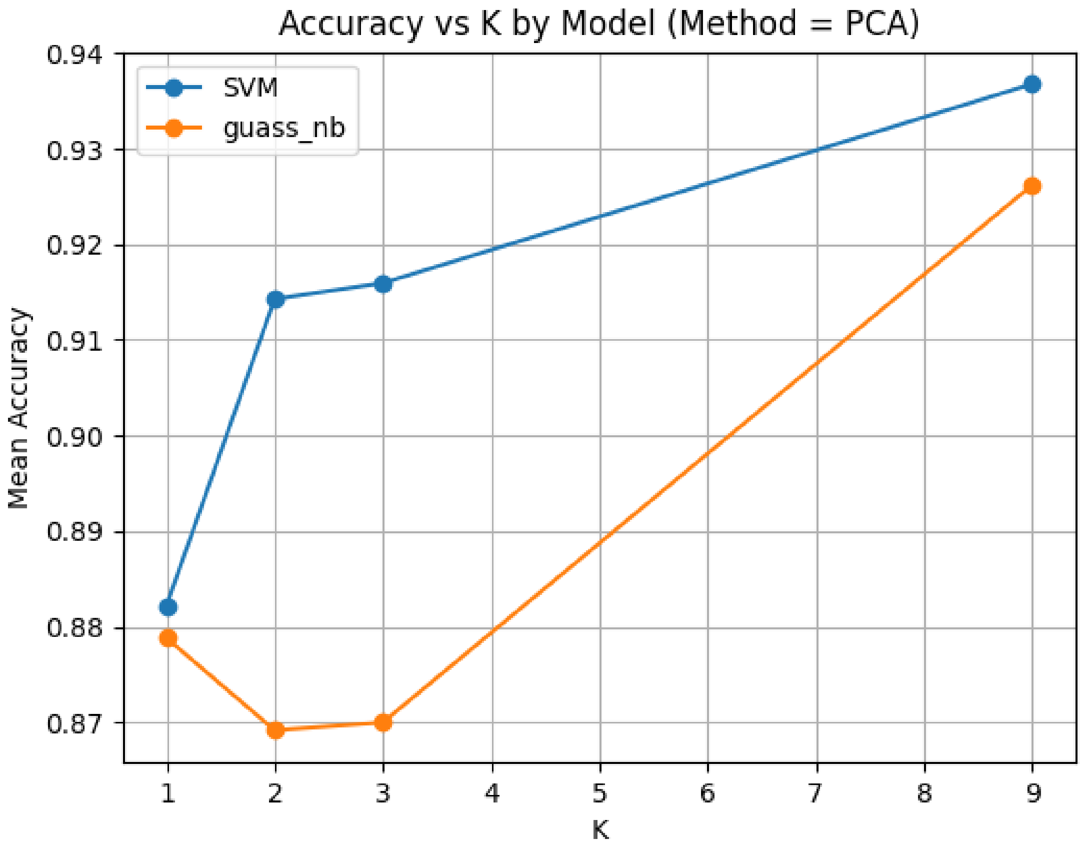


- **a.** SVM and GNB performed very similar for all K with a slight advantage for SVM at higher K.
- **b.** Selecting features via feature importance score using random forest classifier helped model to train well even with very few features (K=1), suggesting that the top-ranked variables are highly informative for the target.
- **c.** Comparing with univariate feature selection method we can infer that using classifier such as Random Forest to compute feature importance is better approach for this type of data.

3. Principle Component Analysis (PCA)

Average accuracy for PCA was also high and relatively stable:

- **K=1:** 0. 880496
- **K=2:** 0. 891728
- **K=3:** 0. 892968
- **K=9:** 0. 931504 (All features)



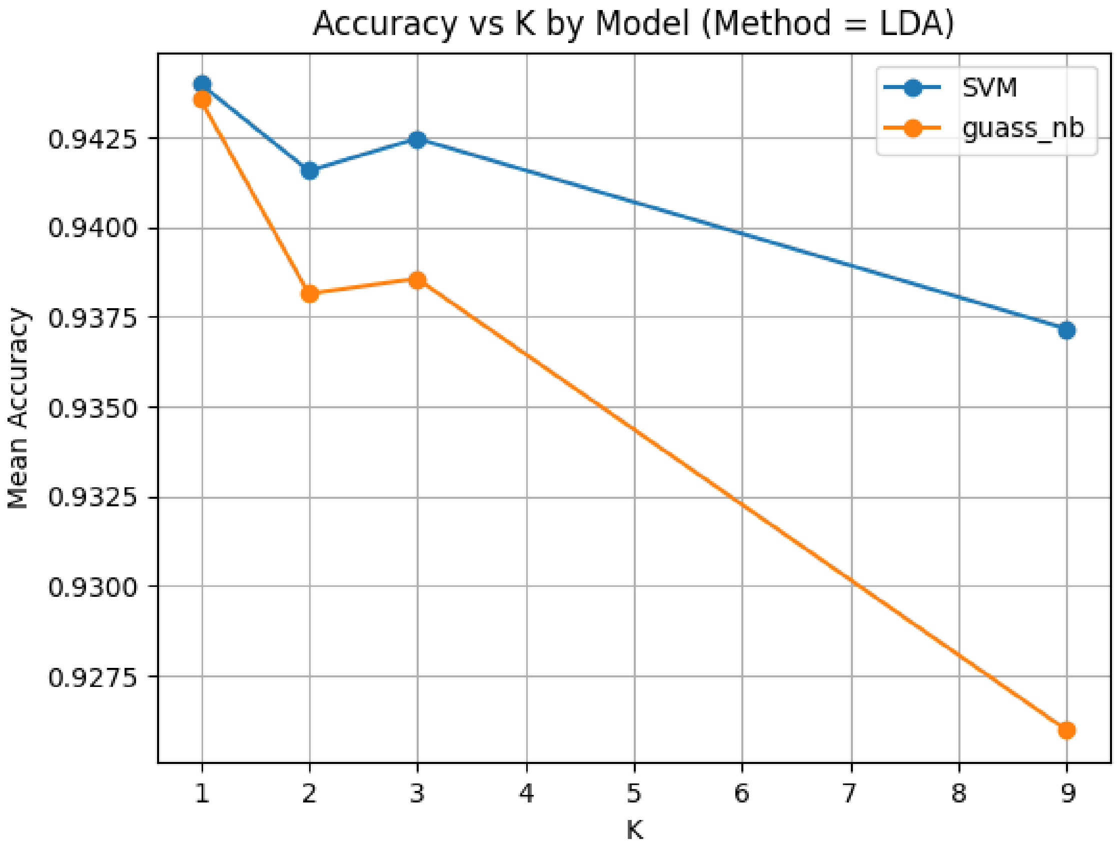
a. SVM clearly outperformed GuassianNB for K=2 and K=3 which indicates that SVM perform better when less dimensions are there for the classification.

b. Although PCA is unsupervised, just using covariance matrix the transformation output generated by PCA still captures structure that is predictive of the classes. However, because PCA optimizes variance rather than class separability, its performance remains slightly below (LDA we will discuss in next section).

4. Linear Discriminant Analysis (LDA)

LDA achieved the highest accuracies overall, and also performed well in lower dimensions :

- **K=1:** 0. 943768
- **K=2:** 0. 939856
- **K=3:** 0. 940512
- **K=9:** 0. 931576 (All features)



a. SVM and GNB performed almost similarly for K=1 to 3, with comparatively a small drop at K=9. Which suggest that when we apply classifier on top of LDA feature selection method it performs well for lower dimensions but fails for relatively higher dimensions.

b. As LDA is supervised learning method of dimensionality reduction it has ability to maximize the separation between-class, it creates a very discriminative low-dimensional space. The fact that K=1 already reaches ≈0.94 accuracy was expected because experience variance ration of PCA1 was **0.98**.

5. Summary:

a. Different feature selection methods, produce different results.

b. It depend on which type of data you are using and what you goal. (Objective)

c. According to objective you may choose appropriate data.

d. From this project I infer that, dimensionality reduction helps to reduce the computational time and cost. Although all features gave us highest accuracy but difference between K=3 and K=9 is not significant.

e. So, we can say that, dimensionality reduction really helps when you are computing large number of features