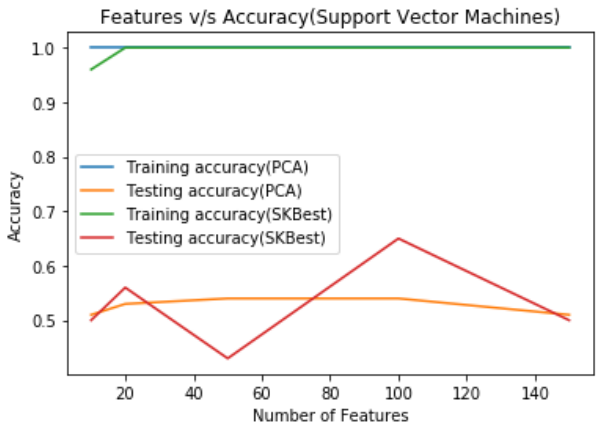
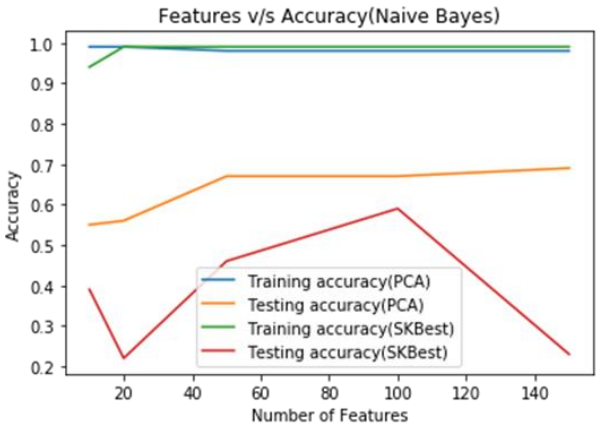
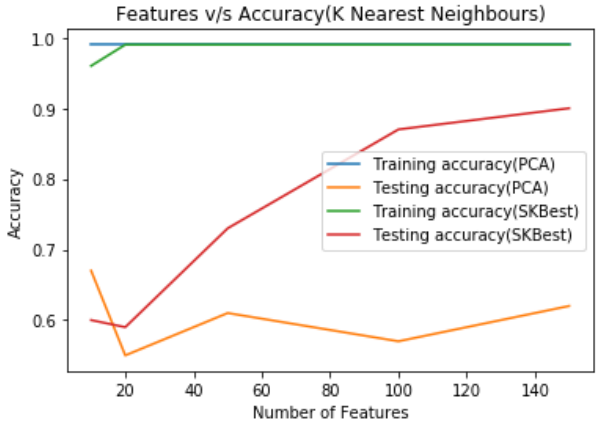
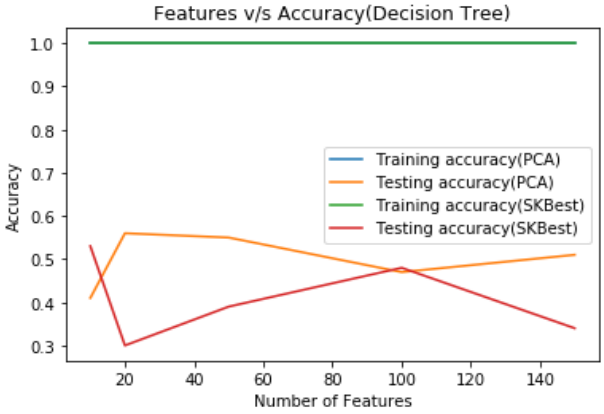
# Feature Analysis

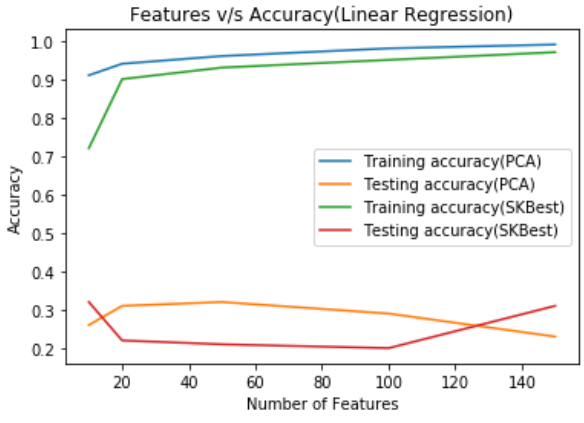
- For each machine learning model, analysis of training and testing accuracy with respect to different feature set.

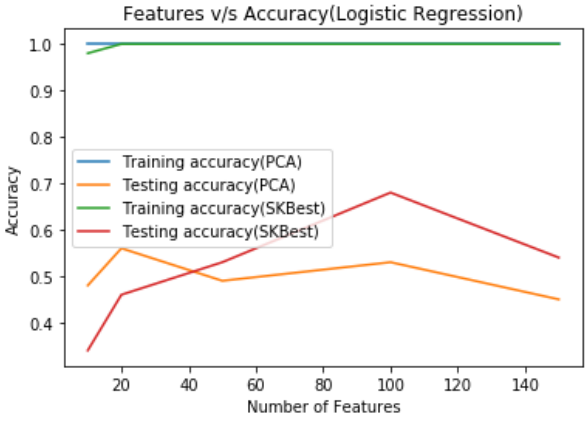


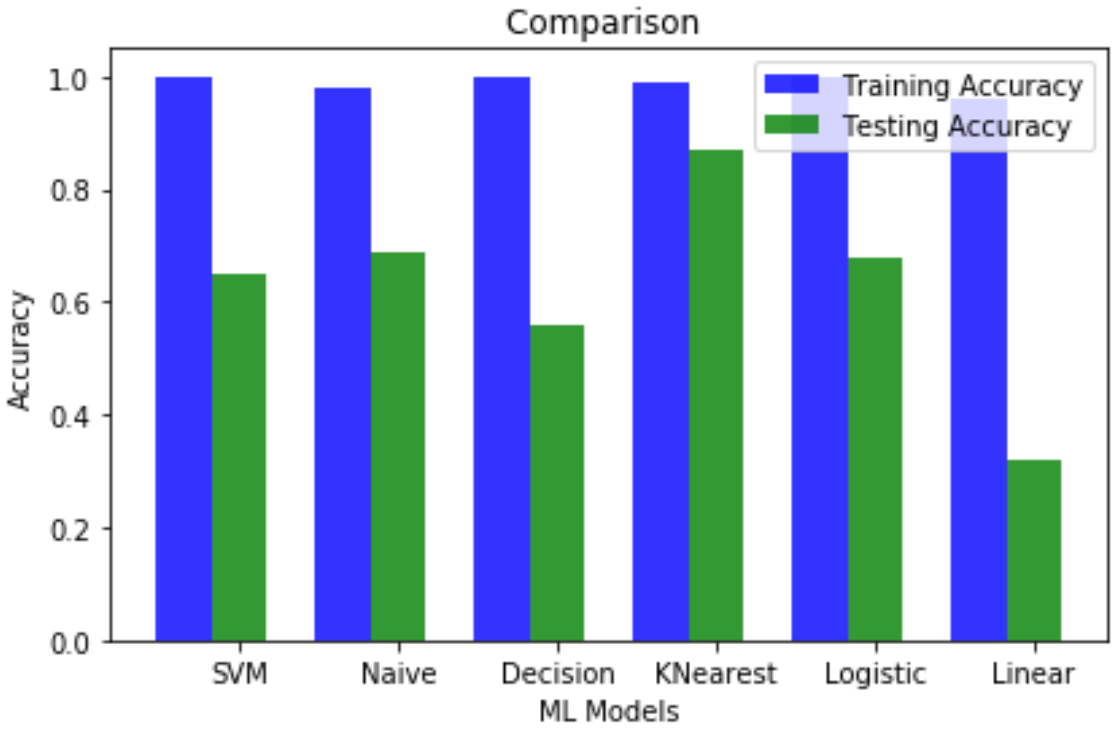



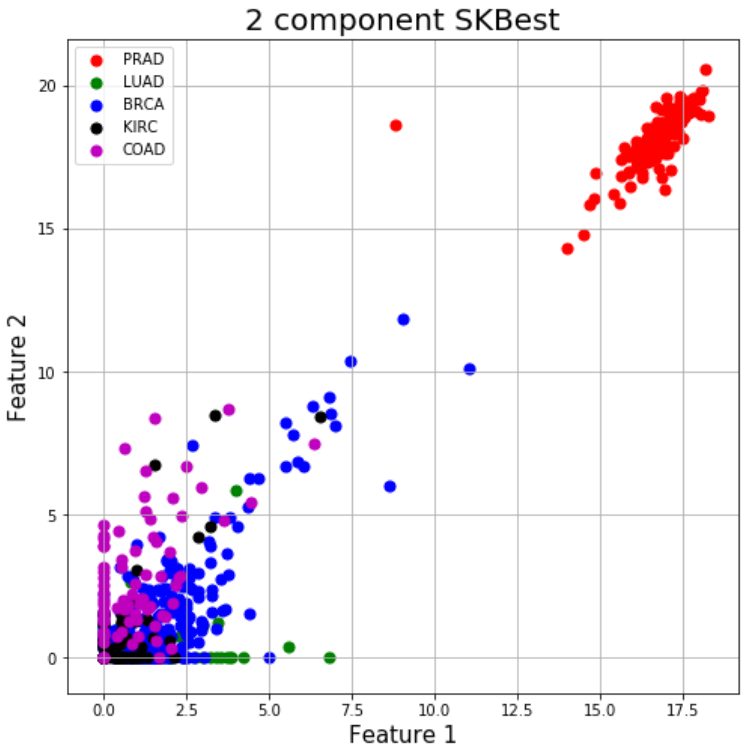



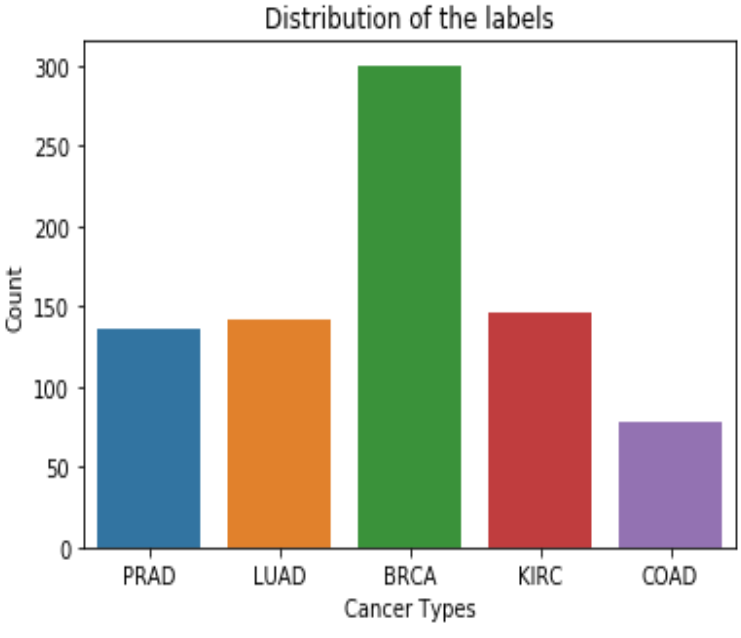



- Comparison among different models by considering the best feature set for each model.




- Distribution of initial data in 2 dimensions



- Classification report and Confusion matrix for k=5 folds (SKBest, K Nearest Neighbours, features=100)

**Fold 1:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 1.00 | 0.98 | 59 |
| 1 | 1.00 | 0.94 | 0.97 | 16 |
| 2 | 1.00 | 0.97 | 0.98 | 29 |
| 3 | 0.96 | 0.93 | 0.95 | 29 |
| 4 | 1.00 | 1.00 | 1.00 | 28 |
| avg / total | 0.98 | 0.98 | 0.98 | 161 |

```
[[59  0  0  0  0]
 [ 0 15  0  1  0]
 [ 1  0 28  0  0]
 [ 2  0  0 27  0]
 [ 0  0  0  0 28]]
```

**Fold 2:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 1.00 | 0.91 | 51 |
| 1 | 1.00 | 1.00 | 1.00 | 20 |
| 2 | 1.00 | 1.00 | 1.00 | 29 |
| 3 | 1.00 | 0.89 | 0.94 | 36 |
| 4 | 1.00 | 0.75 | 0.86 | 24 |
| avg / total | 0.95 | 0.94 | 0.94 | 160 |

```
[[51  0  0  0  0]
 [ 0 20  0  0  0]
 [ 0  0 29  0  0]
 [ 4  0  0 32  0]
 [ 6  0  0  0 18]]
```

**Fold 3:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 1.00 | 0.99 | 60 |
| 1 | 0.82 | 1.00 | 0.90 | 9 |
| 2 | 1.00 | 0.97 | 0.99 | 34 |
| 3 | 1.00 | 1.00 | 1.00 | 28 |
| 4 | 1.00 | 0.93 | 0.96 | 29 |
| avg / total | 0.98 | 0.98 | 0.98 | 160 |

```
[[60  0  0  0  0]
 [ 0  9  0  0  0]
 [ 1  0 33  0  0]
 [ 0  0  0 28  0]
 [ 0  2  0  0 27]]
```

**Fold 4:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 1.00 | 0.89 | 63 |
| 1 | 0.29 | 0.40 | 0.33 | 15 |
| 2 | 1.00 | 1.00 | 1.00 | 26 |
| 3 | 0.77 | 0.92 | 0.84 | 25 |
| 4 | 1.00 | 0.13 | 0.23 | 31 |
| avg / total | 0.82 | 0.76 | 0.72 | 160 |

```
[[63  0  0  0  0]
 [ 2  6  0  7  0]
 [ 0  0 26  0  0]
 [ 2  0  0 23  0]
 [12 15  0  0  4]]
```

**Fold 5:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 1.00 | 0.96 | 67 |
| 1 | 1.00 | 1.00 | 1.00 | 18 |
| 2 | 1.00 | 1.00 | 1.00 | 28 |
| 3 | 1.00 | 0.96 | 0.98 | 23 |
| 4 | 1.00 | 0.79 | 0.88 | 24 |
| avg / total | 0.97 | 0.96 | 0.96 | 160 |

```
[[67  0  0  0  0]
 [ 0 18  0  0  0]
 [ 0  0 28  0  0]
 [ 1  0  0 22  0]
 [ 5  0  0  0 19]]
```

- **Approaches tried to overcome overfitting**
  - Feature selection methods- LDA, PCA and Annova (with 10-200 features)
  - Different classifiers- SVM, Naïve Bayes, Decision Tree and K Nearest Neighbours (with k=5)
  - Cross validation using 3,5,10 folds
  - GridSearchCV for SVM
  - Normalisation of dataset
  - Regularisation
  - Manually creating a balanced subset using original dataset