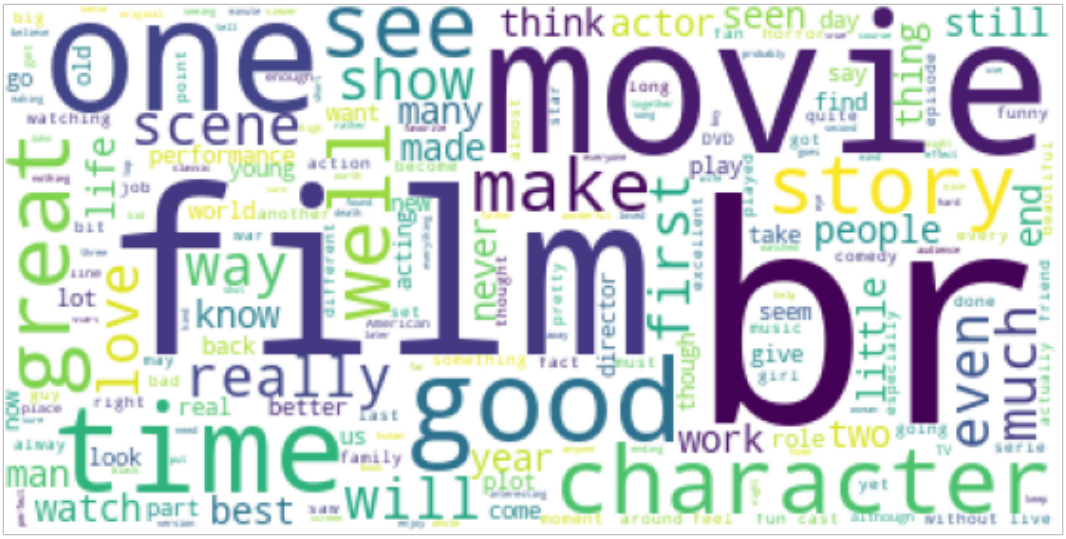
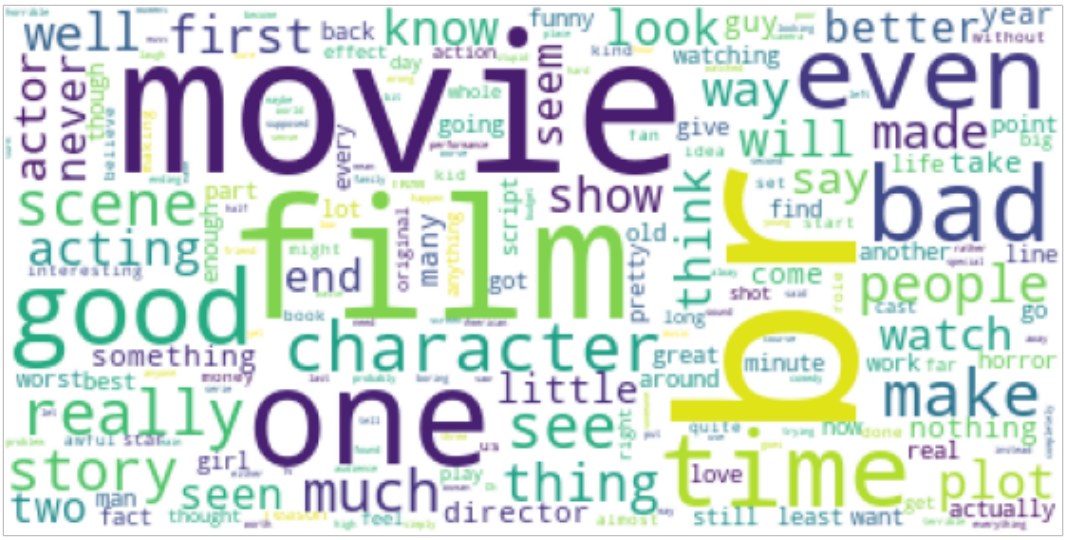
ML Assignment-3 Report

Q1.

1. **Word Cloud For:**
   a. Positive Reviews:



   b. Negative Reviews:

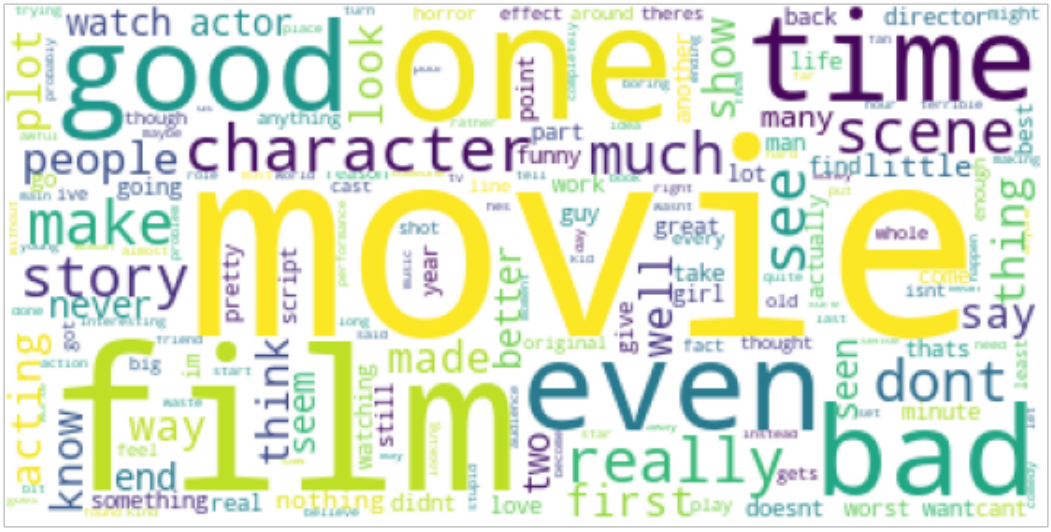


2. **Pre-Processing of text is done by doing following :**

   WordClouds after doing the pre-processing:

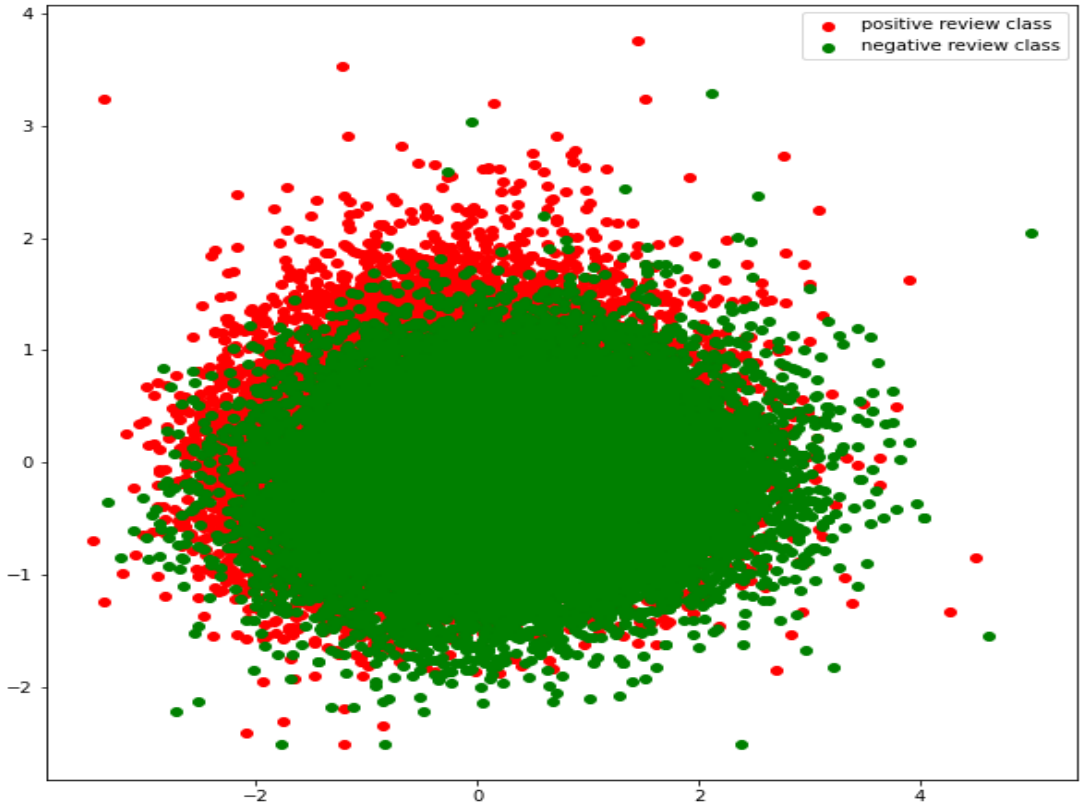
Positive Review:

Negative Review:



Yes, there is a difference in word clouds after pre-processing. Tags like br, stop-words have been removed which were there before pre-processing .

3. **Word2Vec and the PCA plot**



No, classes are not separable in the above PCA plot

4. SVM with different Kernel Observation:

Kernel	Training Accuracy	Testing Accuracy
RBF	0.8685	0.868
Linear	0.85975	0.8621
Poly	0.869975	0.8672
Sigmoid	0.6816	0.6869

RBF Kernel performs better than other kernels. So, it is the best model.

Q2

1. A)PCA :

PCA is called with configuration n_components=0.9, svd_solver='full'. After PCA we got images with 298 dimensions

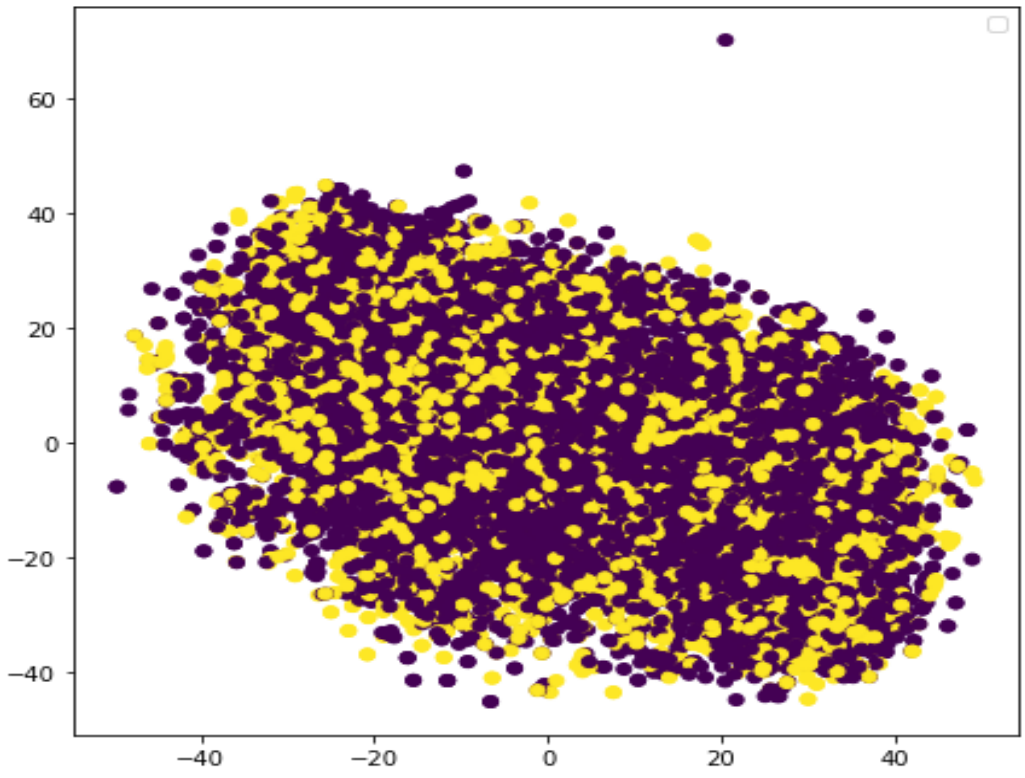
1. The loaded image resize to 32*32. So, every image is of size 32*32*3.
2. After that we flatten every image , every image is feature set of dimension 3072
3. After , PCA every image is of dimension 298

B) Canny Edge Detection and colored histogram:
Canny edge detection is performed as described in the blog post. I have modified the double threshold as per my need. We applied canny-edge detection after converting our images from RGB to greyscale. So, the greyscale images after canny-edge detection is of size 32*32, which we flatten to get a feature descriptor of size 1024

Color Histogram:
To Apply color histogram , I applied it using Wikipedia example which divided every pixel color into 4 bins.
So, there be total 64 possible combination of Red Green and Blue bin numbers. Each of these combination occurences are counted in our image that was our color histogram. So, feature descriptor of color histogram is of size 64 .
So, we got our complete feature descriptor CED+Color Histogram by appending color histogram descriptor with CED feature descriptor. So, this gives us complete feature descriptor of size 1088

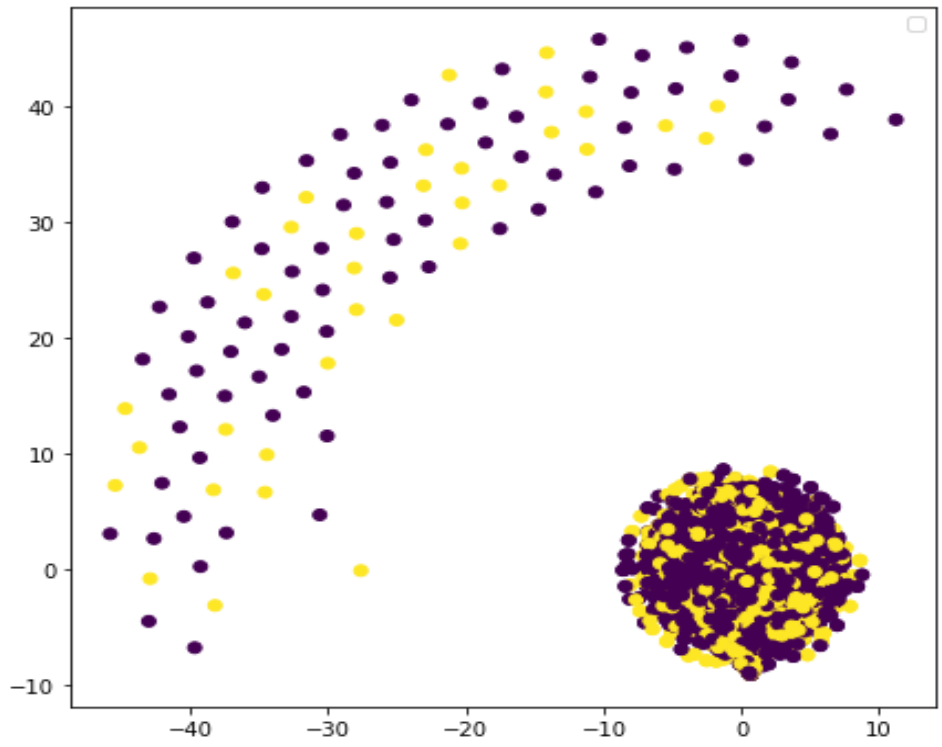
2. 2d TSNE Plot

PCA:
Obeservation: Data is not separable as shown in the above tsne plot

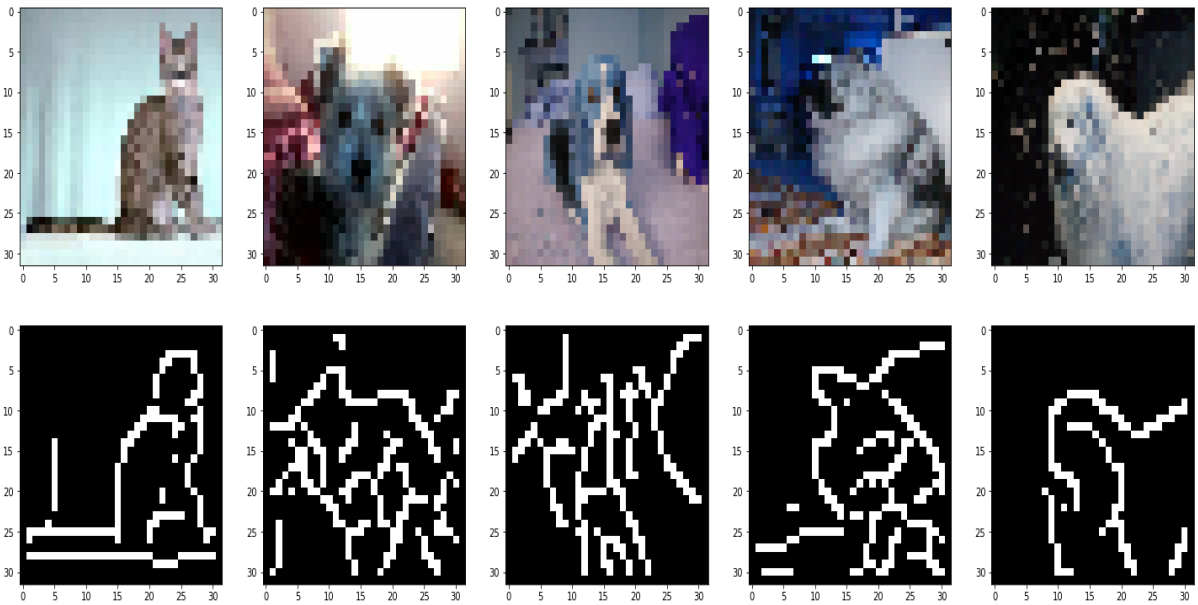


CEd+Histogram

Observation: Data is not separable



Visualisation of 5 images from features extracted using CED



3. **GridSearch CV**

For PCA data
After grid search it was known that the best value for C is 1 with runtime of 8.0s

Training Accuracy: 0.802065695902472

Testing Accuracy: 0.7149627623561273

For Canny edge and histogram data
After grid search it was known that the best value for C is 1 with runtime of 32.0s

Training Accuracy: 0.8838469353200136

Testing Accuracy: 0.6953283683141503

4. New Training data is obtained and following are the observation after svm on new training data

For PCA data
Training Accuracy: 0.7361471861471861
Testing Accuracy: 0.7163168584969533

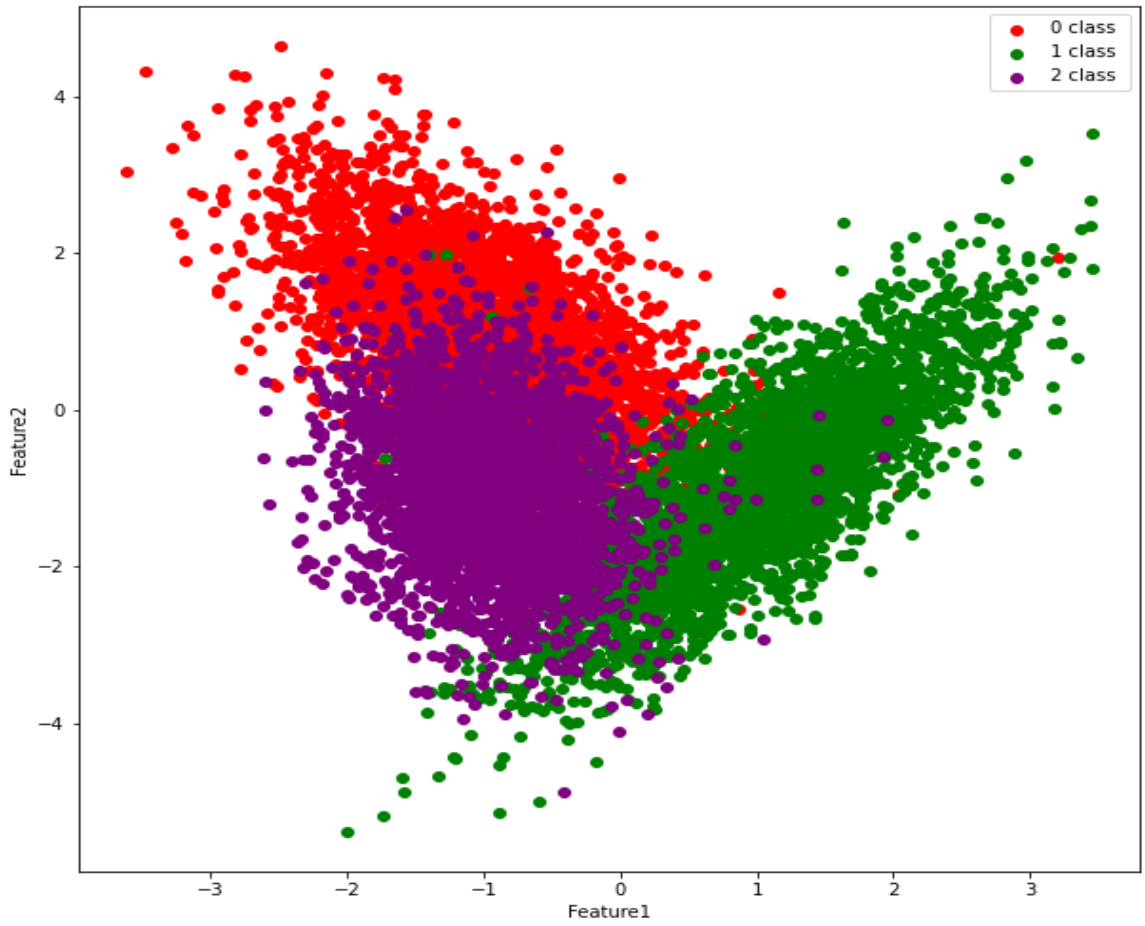
For Canny edge and histogram data
Training Accuracy: 0.8708473310936917
Testing Accuracy: 0.6953283683141503

For PCA data there is slight increase in the training accuracy but testing accuracy is same
For Canny Edge and histogram data there is slight decrease in training accuracy but testing accuracy is same

Q3

Assumption: Data is converted to data frame with samples set as feature1 and feature2 and labels as label

1:
Visualization of the data set :



Observations: Classes are not linearly separable

Some few observations on data are as follows:

a. Classwise distributation of data:

0 class: 3341
1 class: 3332
2 class: 3327

b. Data info:

```
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
0   feature1  10000 non-null  float64
1   feature2  10000 non-null  float64
2   label     10000 non-null  int64
```

c. Null values in data:

```
feature1    0
feature2    0
label       0
```

d. NAN values in data:

```
feature1    0
feature2    0
label       0
```

e. Description of data:

	feature1	feature2	label
count	10000.000000	10000.000000	10000.000000
mean	-1646.361257	-1741.602051	5003.431800
std	954.045444	997.686086	2888.629053
min	-3318.940355	-3448.540498	1.000000
25%	-2444.825021	-2604.657711	2486.000000
50%	-1649.280309	-1761.008233	5013.000000
75%	-814.452303	-855.425749	7525.250000
max	0.247028	-2.536864	9986.000000

2: OneVsRest Approach:

Grid Search on C on 1 fold is as follows:

Accuracy for C value 1 0.8765
Accuracy for C value 11 0.876875
Accuracy for C value 21 0.877375
Accuracy for C value 31 0.878
Accuracy for C value 41 0.878
Accuracy for C value 51 0.877875
Accuracy for C value 61 0.87775
Accuracy for C value 71 0.87775
Accuracy for C value 81 0.877875
Accuracy for C value 91 0.877875

After grid search optimal value for C obtained is 41 with accuracy 0.878

Accuracies of 5-folds are as follows:

Folds	Training Accuracy	Testing Accuracy
Fold-1	0.873375	0.8835
Fold-2	0.877	0.871
Fold-3	0.875875	0.8765
Fold-4	0.87725	0.8735
Fold-5	0.877875	0.871

Mean Foldwise accuracies are:
Training mean accuracy : 0.876275
Testing mean accuracy : 0.8751

Classwise Accuracy

Folds	Class 0	Class 1	Class 2
Fold-1	0.85864	0.9555	0.8348
Fold-2	0.862	0.9496	0.8038
Fold-3	0.8618	0.9568	0.8114

Fold-4	0.8602	0.9528	0.8076
Fold-5	0.8539	0.9508	0.8061

Mean Class Accuracy:

Class 0: 0.85935
Class 1: 0.95315
Class 2: 0.812774

3:OneVsOne Approach:

Grid Search on C on 1 fold is as follows:

Accuracy for C value 1 0.877875
Accuracy for C value 11 0.8775
Accuracy for C value 21 0.8775
Accuracy for C value 31 0.877375
Accuracy for C value 41 0.87725
Accuracy for C value 51 0.877
Accuracy for C value 61 0.87725
Accuracy for C value 71 0.877125
Accuracy for C value 81 0.877125
Accuracy for C value 91 0.877

After grid search optimal value for C obtained is 21 with accuracy 0.8775

Accuracies of 5-folds are as follows:

Folds	Training Accuracy	Testing Accuracy
Fold-1	0.873	0.887
Fold-2	0.87675	0.8715
Fold-3	0.8755	0.8785
Fold-4	0.87875	0.872
Fold-5	0.877	0.8715

Mean Foldwise accuracies are:

Training mean accuracy : 0.8762

Testing mean accuracy : 0.87609

Classwise Accuracy

Folds	Class 0	Class 1	Class 2
Fold-1	0.8646	0.9585	0.8363
Fold-2	0.8635	0.9496	0.8038
Fold-3	0.8695	0.9568	0.8099
Fold-4	0.8588	0.9513	0.8061

Fold-5	0.8584	0.9494	0.8046

Mean Class Accuracy:

Class 0: 0.863001
Class 1: 0.953150
Class 2: 0.812172

4: Sklearn implementation

OneVsRest Accuracies on test set foldwise:

Folds	Test Accuracy
Fold-1	0.8865
Fold-2	0.871
Fold-3	0.879
Fold-4	0.8725
Fold-5	0.8715

Test Mean Accuracy: 0.87609

OneVsOne Accuracies on test set foldwise:

Folds	Test Accuracy
Fold-1	0.8865
Fold-2	0.871
Fold-3	0.879
Fold-4	0.8725
Fold-5	0.8715

Test Mean Accuracy: 0.87609

Yes , there is slight deviation in OneVsRest scratch and sklearn model performance but there is no deviation in OneVsOne sratch and sklearn model performance
Deviations are as shown in table below:

Model	Mean Test Accuracy
OneVsRest Scratch	0.8751
OneVsRest Sklearn	0.87609
OneVsOne Scratch	0.87609
OneVsOne Sklearn	0.87609