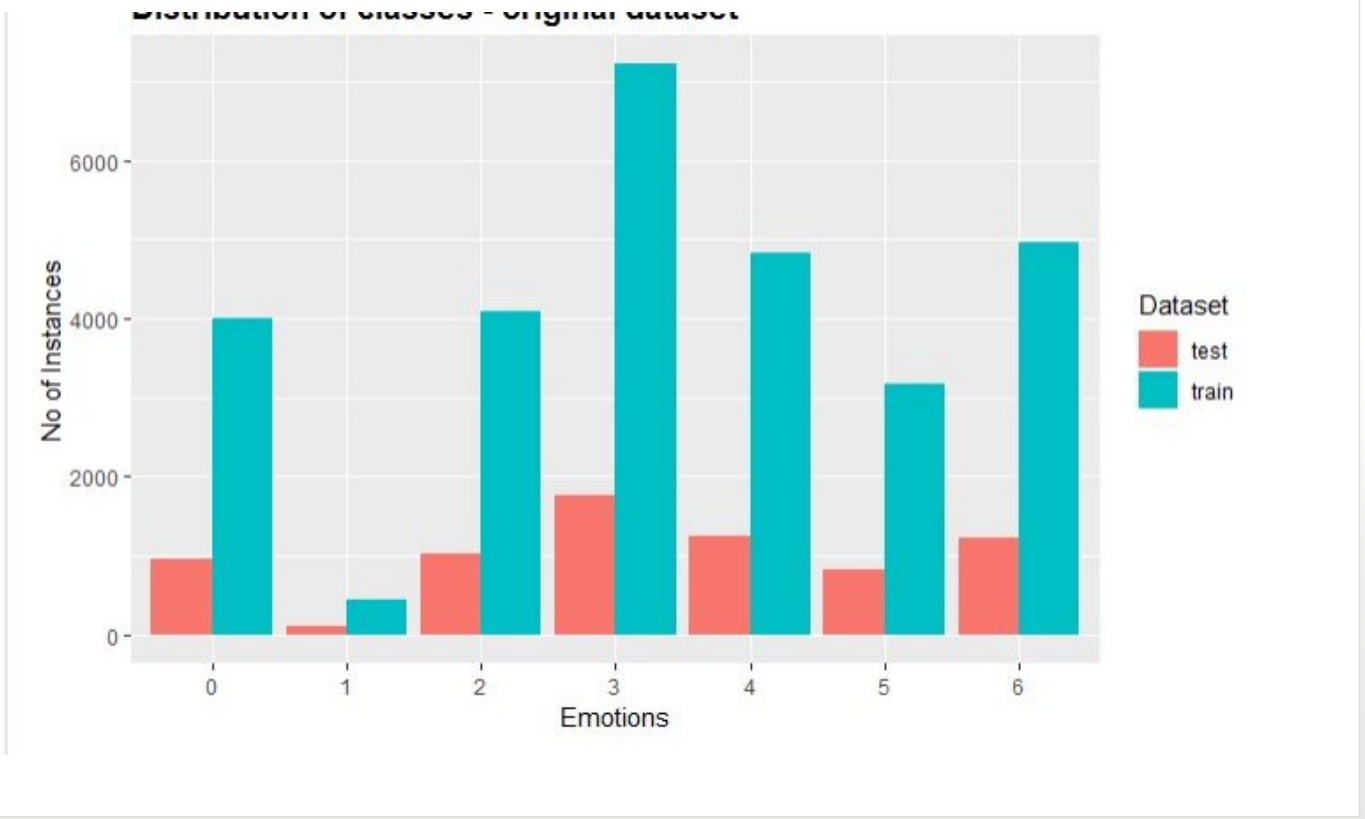
STRATIFICATION & DATA PREPARATION

NISHNA AJMAL

INPUT

- Facial Image Recognition Dataset
  - Contains 36000 images (48 x 48 pixels)of faces almost centered so that all the faces takes up nearly equal spaces in the image
- ML Algorithm : Decision Trees
- Model Programmed in R
  - rpart , caret – DT
  - matrixstats – summary statistics
  - ggplot – visualizations
  - EBImage – Image processing

DISTRIBUTION OF CLASSES IN THE DATASET



Nishna

	cp <dbl>	Accuracy <dbl>	Kappa <dbl>	AccuracySD <dbl>	KappaSD <dbl>
1	0.001128222	0.2665855	0.04763352	0.006025291	0.01173029
2	0.001170869	0.2660456	0.04588583	0.005939549	0.01134324
3	0.003908067	0.2541501	0.01176409	0.004282721	0.01382140

3 rows

DECISION TREE CLASSIFIER PERFORMANCE – BEFORE PREPROCESSING

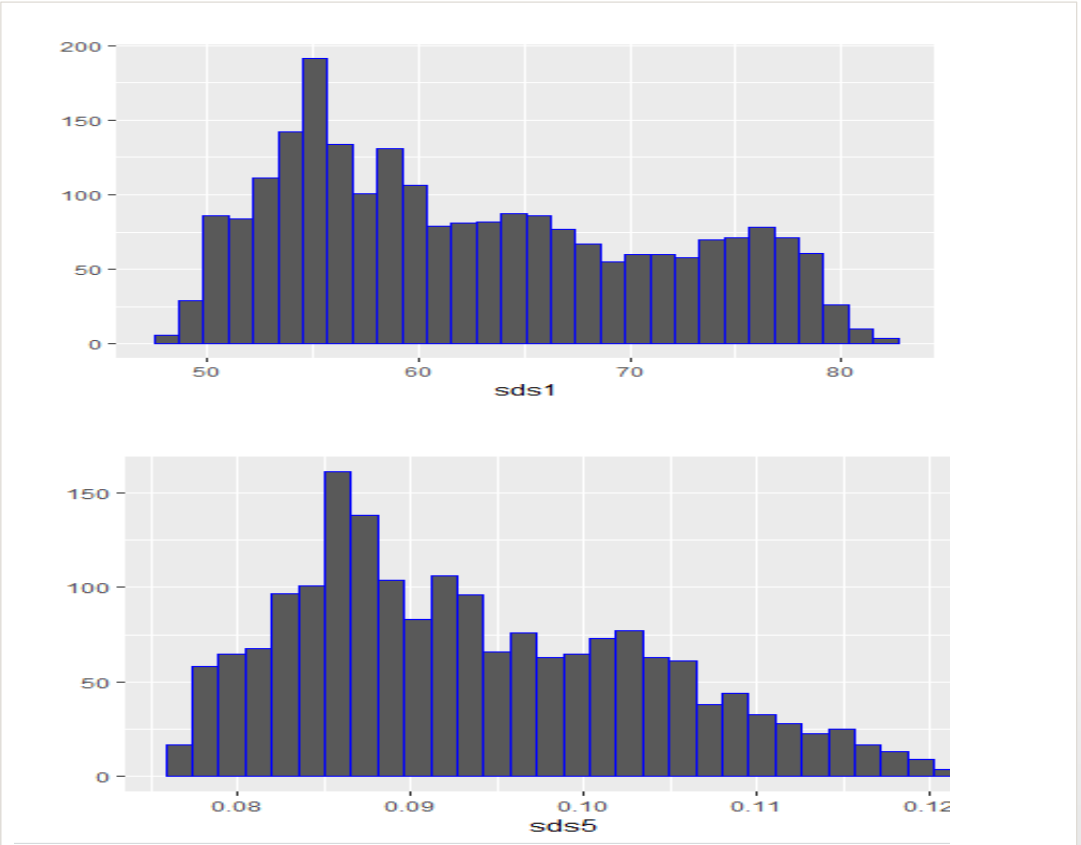
DATA PREPARATION

- Rotation
- Cropping
- Centering
- Scaling
- PCA
- SMOTE
- Under sampling

SAMPLE IMAGE

ROTATED & CROPPED

Nishma

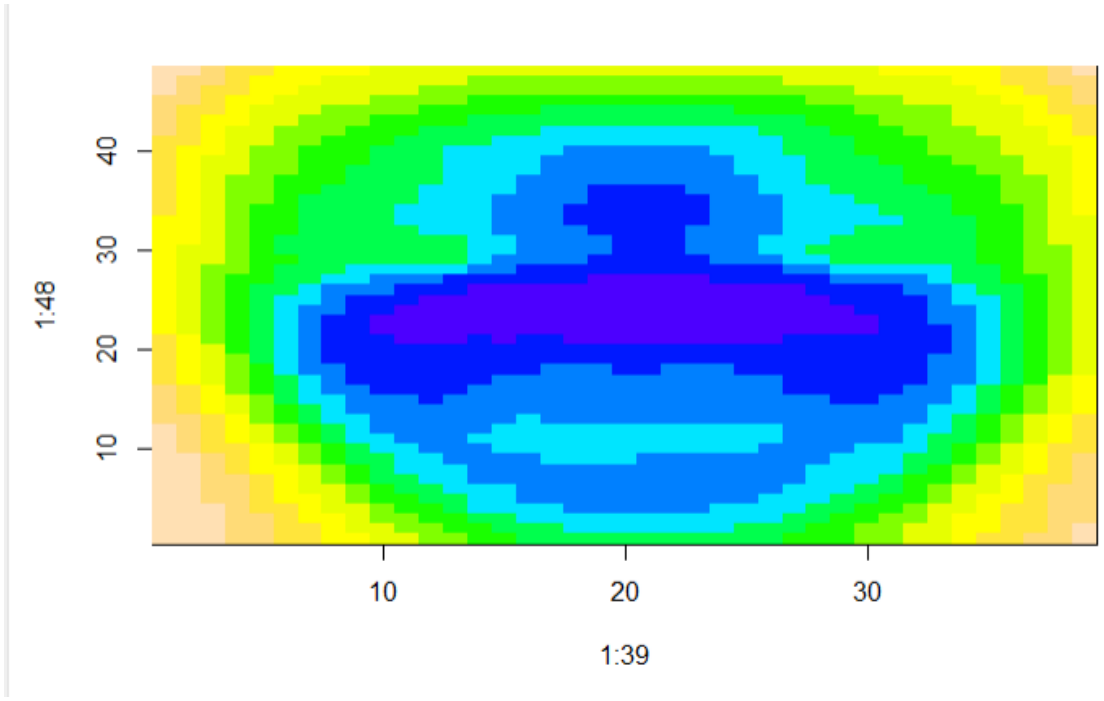


DISTRIBUTION OF COLUMN STANDARD DEVIATION

- Fig I : Original dataset
- Fig II : Rotated, cropped, centered & scaled

Nishna

DISTRIBUTION OF VALUES IN EACH PIXEL

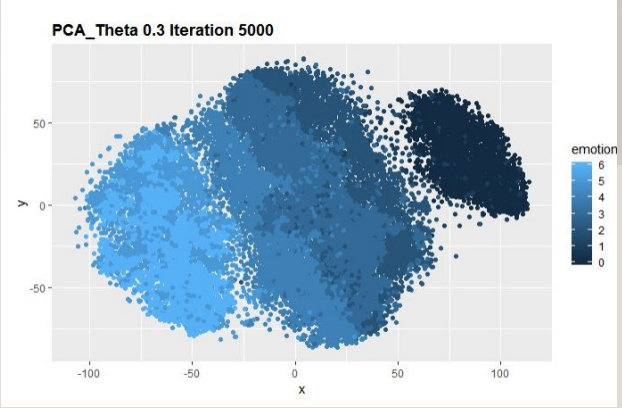


	PC2	PC5	PC10	PC15	PC20	PC25	PC50
Standard deviation	1.883853	1.017172	0.5280574	0.3813874	0.3211437	0.2855511	0.1754931
Proportion of Variance	0.178620	0.052070	0.0140300	0.0073200	0.0051900	0.0041000	0.0015500
Cumulative Proportion	0.426680	0.621580	0.7128500	0.7594800	0.7899800	0.8115300	0.8721400

PCA SUMMARY

- First 25 components holds more than 80% of the data

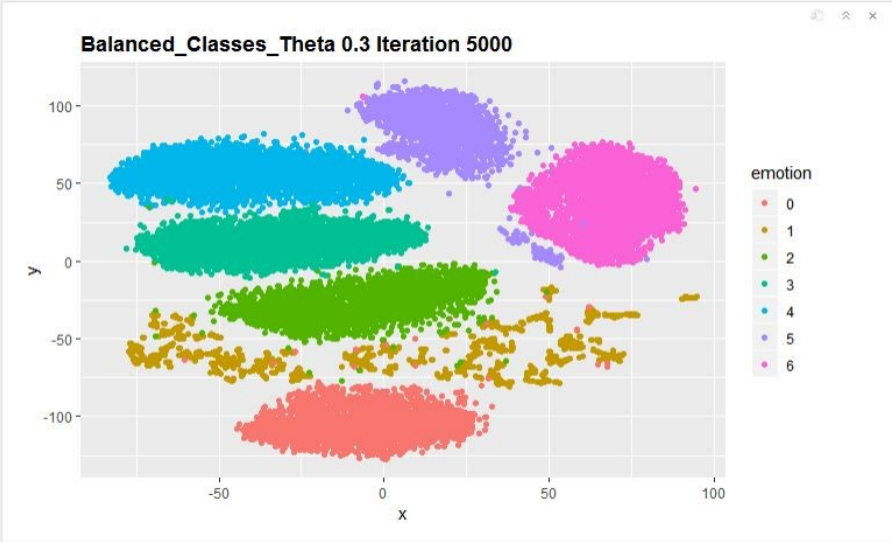
Nishna



T-SNE VISUALIZATION

- The first fig shows the original dataset with 48 x 48 = 2304 features visualized in 2D.
- The data which after preprocessing(rotated, cropped, scaled, centered) and PCA(reduced dimensions to 25) can be visualized to 2D as shown in fig 2.

Nishna



T-SNE VISUALIZATION - II

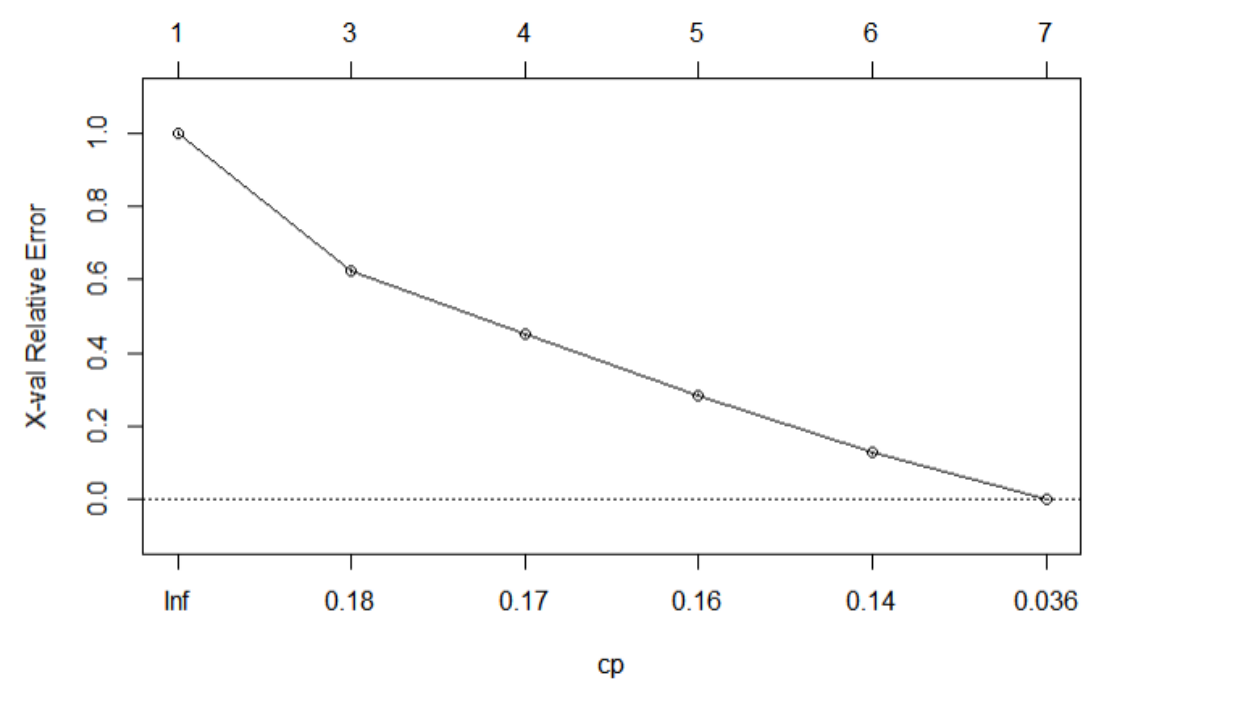
- The prepared training dataset is then balanced by
  - Synthetic Oversampling using SMOTE
  - Randomly Undersampling the minority class

CLASSIFIER PERFORMANCE

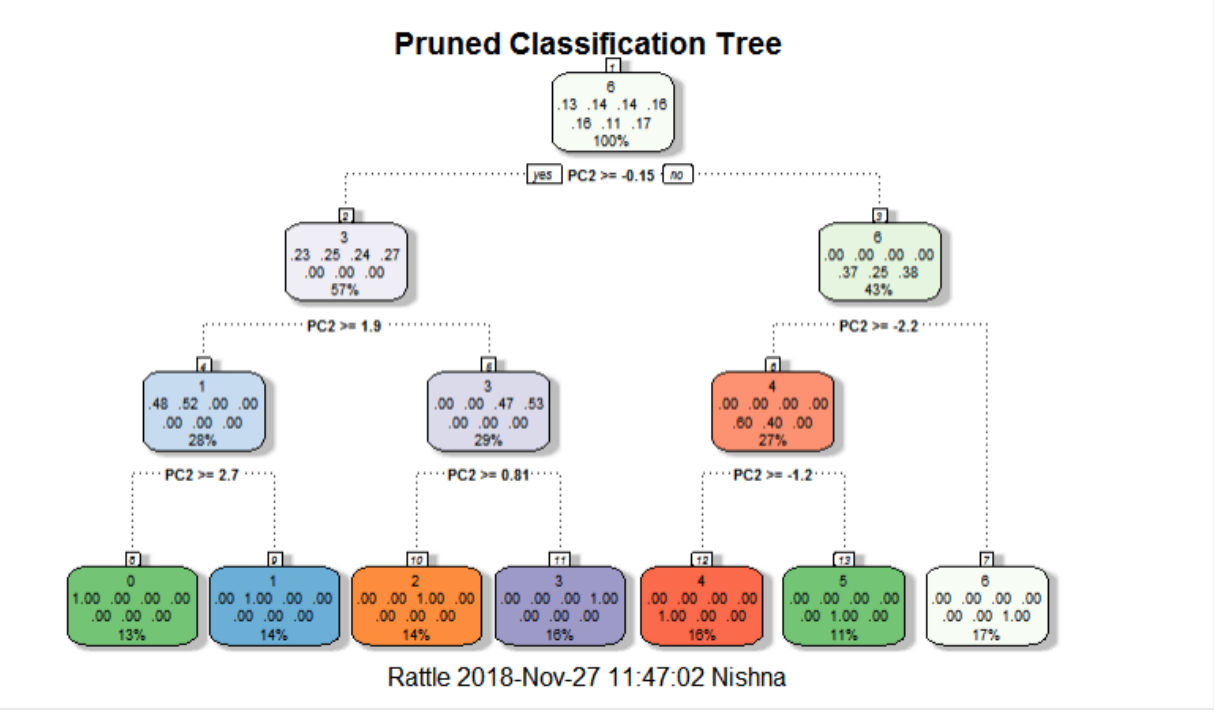
Before hyper parameter tuning

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6
         0    0    0    0    0    0    0    0
         1 1013  110    0    0    0    0    0
         2    0    0    0    0    0    0    0
         3    0    0 1009 1804    0    0    0
         4    0    0    0    0 1222  779    0
         5    0    0    0    0    0    0    0
         6    0    0    0    0    0    0 1242

Overall Statistics

               Accuracy : 0.6098
                 95% CI : (0.5984, 0.6211)
    No Information Rate : 0.2513
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5252
 Mcnemar's Test P-Value : NA
```

HYPER PARAMETER TUNING – COMPLEXITY PARAMETER(CP)



HYPER PARAMETER TUNING – PRUNING

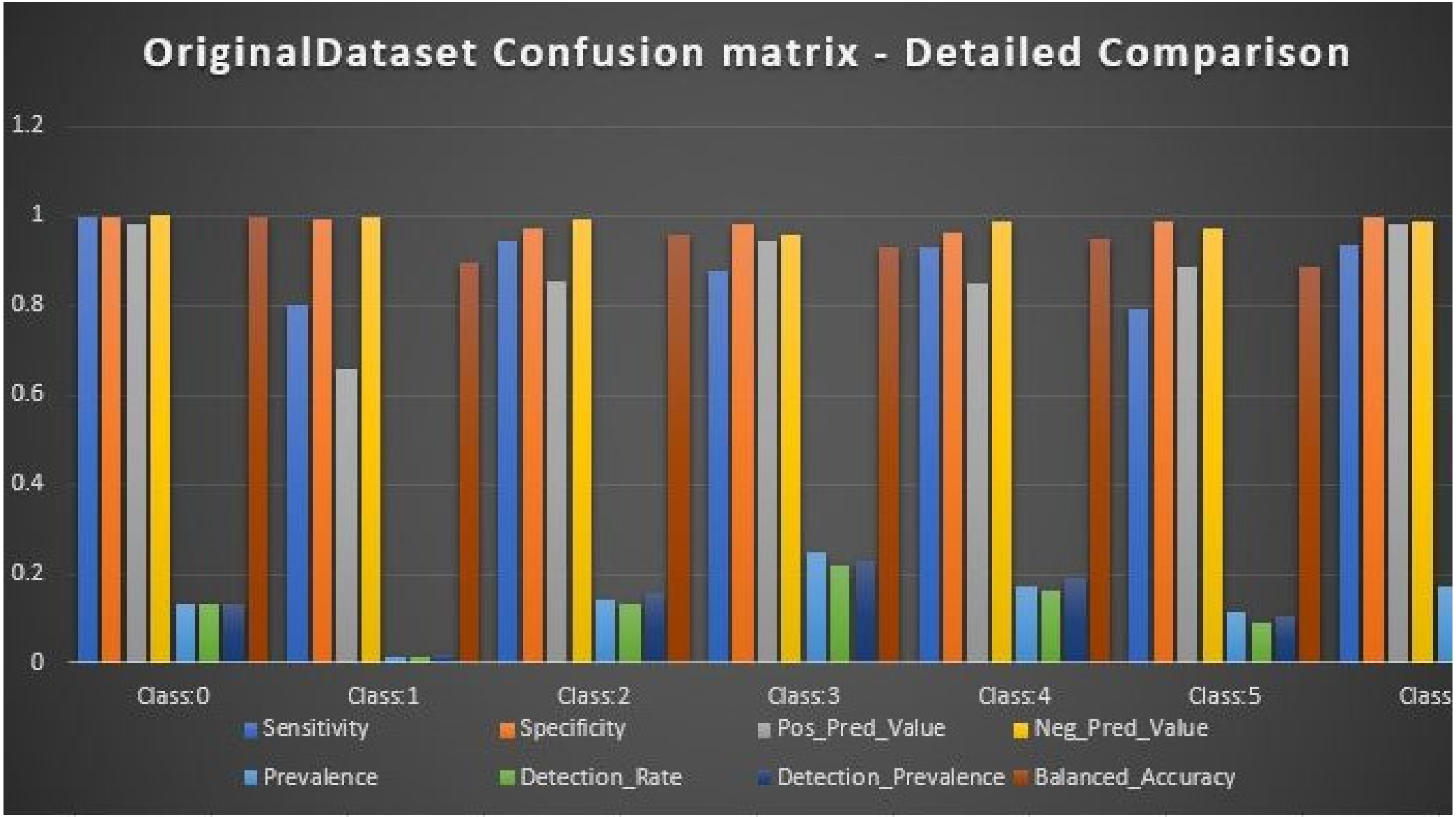


CLASSIFIER PERFORMANCE

After hyper parameter tuning

```
[1] "                CONFUSION MATRIX-FINAL MODEL-ORIGINAL TEST"
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6
         0  954   15    0    0    0    0    0
         1    4   89   42    0    0    0    0
         2    0    7  969  162    0    0    0
         3    0    0   13 1557   82    0    0
         4    0    0    0   55 1160  155    0
         5    0    0    0    0    5  658   80
         6    0    0    0    0    0   18 1153

Overall Statistics

               Accuracy : 0.9111
                 95% CI : (0.9043, 0.9176)
    No Information Rate : 0.2471
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8928
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
Sensitivity            0.9958  0.80180   0.9463   0.8777   0.9302  0.79182   0.9351
Specificity            0.9976  0.99349   0.9725   0.9824   0.9646  0.98661   0.9970
```

Nishna


OriginalDataset Confusion matrix - Detailed Comparison
1.2
1
0.8
0.6
0.4
0.2
0
Class:0
Class:1
Class:2
Class:3
Class:4
Class:5
Class
Sensitivity
Specificity
Pos_Pred_Value
Neg_Pred_Value
Prevalence
Detection_Rate
Detection_Prevalence
Balanced_Accuracy