PROBLEM SETTING:

The major point that we are addressing in this project is to help users improve their state of health by intimating medical practitioner their state of stress or increased stress. An analysis of federal health data found that 8.3 million American adults, or roughly 3.4 percent of the population, experience severe psychological discomfort. The researchers noted that 3 percent or less of Americans were thought to be experiencing substantial psychological discomfort in previous estimations. Hence, in this project we are trying to reduce their mental stress and improving their Mental Health.

PROBLEM DEFINITION:

In this project, we use WESAD which is publicly available dataset for Wearable Stress and Affect Detection. This multimodal dataset features physiological and motion data, recorded from both a wrist- and a chest-worn device, of 15 subjects during a lab study. The following sensor modalities are included: blood volume pulse, electrocardiogram, electrodermal activity, electromyogram, respiration, body temperature, and three-axis acceleration. The goal of the project is to analyze the data of the population, and to predict whether the person is feeling stressed or not within a measuring range of five different levels.

DATA SOURCE CITATION:

This dataset was procured from UCI Machine Learning Repository

Data Source Citation:
https://archive.ics.uci.edu/ml/datasets/WESAD+%28Wearable+Stress+and+Affect+Detection%29

DATA DESCRIPTION:

There are 63,000,000 instances and 12 attributes in the training data. The target variable is 'label' which is an ID of respective study protocol condition. The following IDs are provided: 0 = Transient, 1 = Baseline, 2 = Stress, 3 = Amusement, 4 = Meditation. The data is a multivariate time-series with real-valued attributes. The preferred data operations associated with the dataset are classification and regression.

DATA EXPLORATION:

The data is in pickle format (.pkl) To be able to perform operations on the data we use the 'pickle' package to convert the data from pickle to dictionary format. The dictionary contains 3 main entities -

Subject: The dataset contains of 15 patients. 'Subject' notates the member's medical record that the dataset belongs to.

Signal: It describes the body part of which the stress is being calculated from various attributes. Signal is being calculated from

a. Chest : It has further medical measurements which are used for Stress detection
    i. ACC
    ii. ECG
    iii. EMG
    iv. EDA
    v. Temp
    vi. Resp

b. Wrist: It has some other medical measurements of the wrist included used for Stress detection
    i. ACC
    ii. BVP
    iii. EDA
    iv. Temp

ACC is the accelerometer reading, which are being considered in triaxial orthogonal directions. The triaxial ACC readings are different for chest and wrist. The notations are different to avoid ambiguity, i.e., for

ACC for Chest : c_ax, c_ay, c_az

ACC for wrist: w_ax, w_ay, w_az

Label: The label denotes the target variable for the data mining operation being performed on the dataset. It indicates the state of the member in accordance with the stress level the member is in, for the reading. Each label has a number code. The Labels are:

i. 0 – Baseline

ii. 1 – Stress

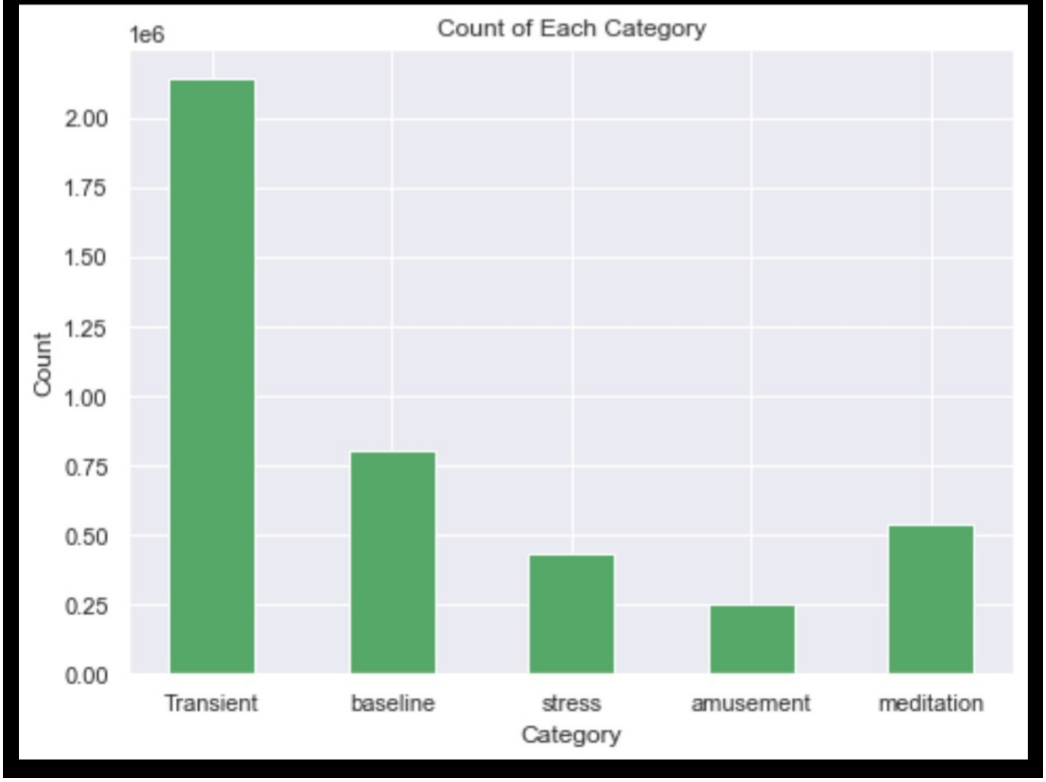
iii. 2 – Amusement

iv. 3 – Meditation

LABELS DISTRIBUTION IN THE DATA:

Below is the data distribution of the counts of labels of the first member

```
transient_indices 2142701
baseline_indices 800800
stress_indices 430500
amusement_indices 253400
meditation_indices 537599
```

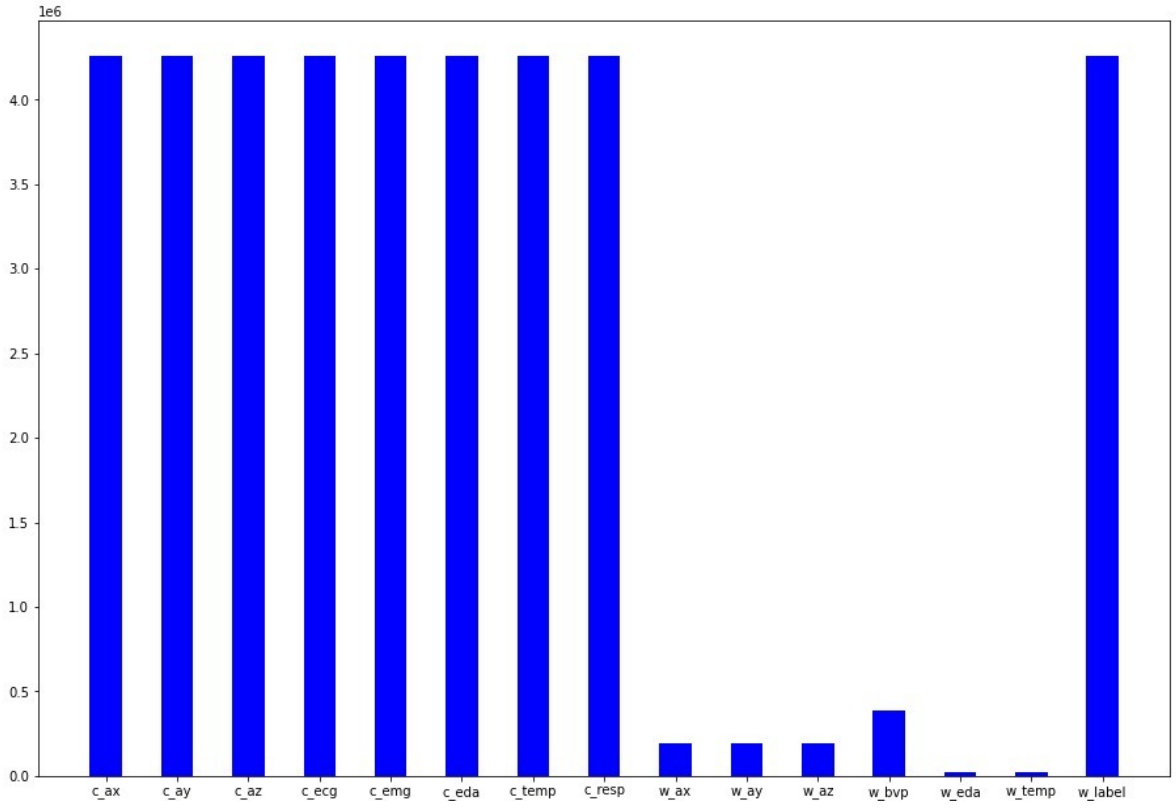


BALANCING UNBALANCED CHEST DATA:

The data samples from chest device are 21 times more than data wrist device samples causing an imbalance, due to which the wrist samples must be excluded from the dataset. Chest device gives 4255300 samples

Below is the distribution visualization of the unbalanced data of chest and wrist devices:

```
c_ax 4255300
c_ay 4255300
c_az 4255300
c_ecg 4255300
c_emg 4255300
c_eda 4255300
c_temp 4255300
c_resp 4255300
w_ax 194528
w_ay 194528
w_az 194528
w_bvp 389056
w_eda 24316
w_temp 24316
[0 0 0 ... 0 0 0] 4255300
Min label value 0 Max label value 7
```



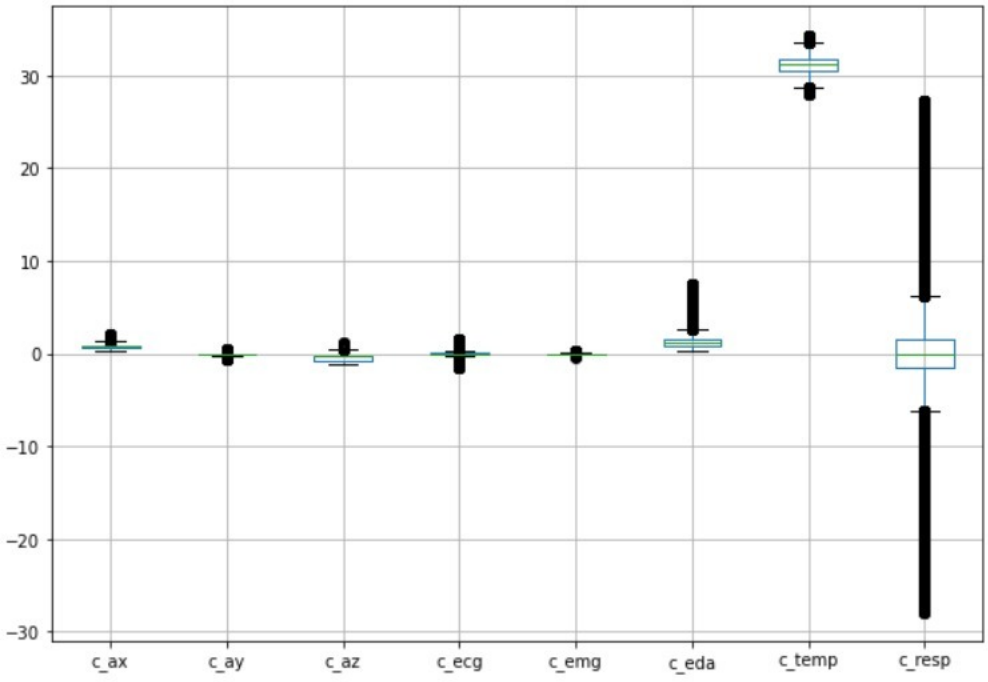
Hence, we convert all the Pickle dictionary records to Dataframe using Pandas by removing wrist samples. Now we only have chest device data records.

Once the data is converted to Dataframes we have the following columns for the data:

“c_ax", "c_ay", "c_az","c_ecg","c_emg","c_eda","c_temp","c_resp","w_label"

CALCULATING INTERQUARTILE RANGE & REMOVING OUTLIERS:

We visualize the data below, by using boxplot:



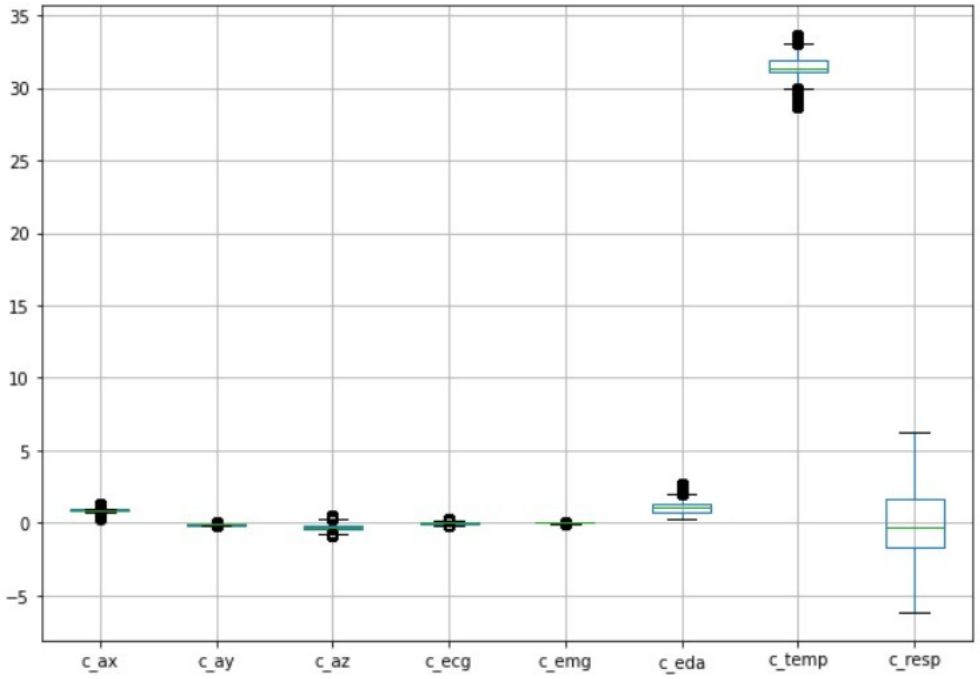
In descriptive statistics, the interquartile range (IQR) is the measure of the spread of the middle half of the data distribution. It is measure of where the bulk of the values of the data lies.

IQR = Q3 – Q1 , where IQR = Interquartile Range

Q3 = 3rd Quartile or the 75th Percentile

Q1 = 1st Quartile or the 25th Percentile

After performing IQR and removing all the outliers, the data is scaled down to a measurable range. The boxplot visualization is as below:



CORRELATION MATRIX:

Below is the correlation matrix of the data:

	c_ax	c_ay	c_az	c_ecg	c_emg	c_eda	c_temp	c_resp	w_label
c_ax	1.000000	-0.023870	0.890463	-0.009258	-0.001814	0.086628	0.121247	-0.018757	-0.509452
c_ay	-0.023870	1.000000	0.027892	0.004532	-0.001511	-0.023619	-0.063603	-0.016261	-0.228733
c_az	0.890463	0.027892	1.000000	-0.004805	-0.002277	-0.138557	0.236908	0.001604	-0.410491
c_ecg	-0.009258	0.004532	-0.004805	1.000000	-0.005601	-0.022128	0.014425	0.064024	0.003692
c_emg	-0.001814	-0.001511	-0.002277	-0.005601	1.000000	-0.005357	-0.003173	-0.000339	-0.006690
c_eda	0.086628	-0.023619	-0.138557	-0.022128	-0.005357	1.000000	-0.541162	-0.030029	-0.119094
c_temp	0.121247	-0.063603	0.236908	0.014425	-0.003173	-0.541162	1.000000	0.019251	0.172247
c_resp	-0.018757	-0.016261	0.001604	0.064024	-0.000339	-0.030029	0.019251	1.000000	-0.003661
w_label	-0.509452	-0.228733	-0.410491	0.003692	-0.006690	-0.119094	0.172247	-0.003661	1.000000

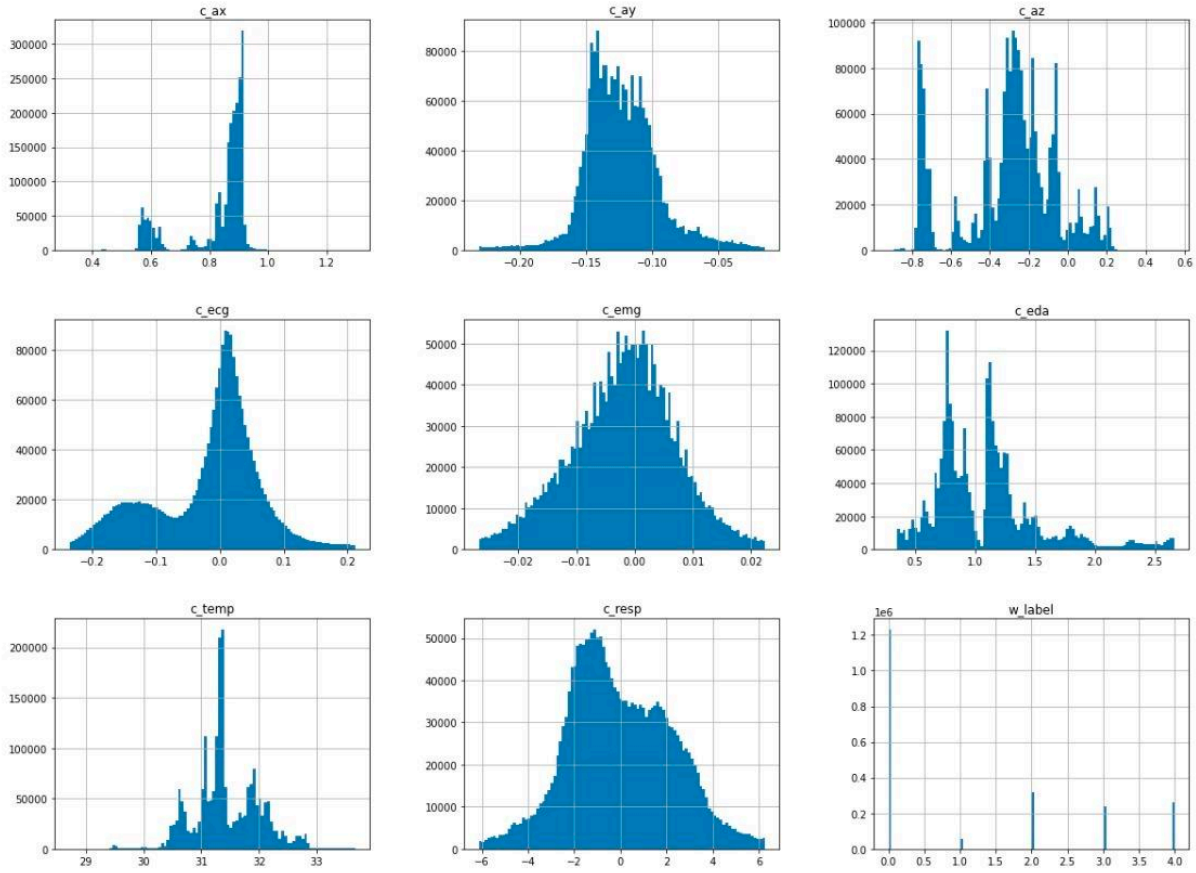
Heatmap of the data:

Interpretations: From the above correlation matrix and heatmap of the data, the following interpretations can be made:

- c_ax is highly positively correlated with c_az
- c_eda and c_temp are negatively correlated with each other
- c_emg and c_ecg are very poorly correlated with the rest of the features.

HISTOGRAM:

Below are the distribution of the data:



Interpretations:

We observe that

- The IQR is the highest for c_emg & c_ay histograms, followed by c_resp and c_ecg graphs
- There is a high varied distribution for the c_az histogram
- A bimodal graph is observed with both c_ax, c_eda histograms
- c_emg shows a uniform distribution graph

DIMENSION REDUCTION AND VARIABLE SELECTION

NORMALIZATION

Z-score normalization is a strategy of normalizing data that avoids this outlier issue. The formula for Z-score normalization is:

$$(\text{Value}-\mu)/\sigma$$

Here, the mean value of the feature is notated by μ, and the standard deviation of the feature is notated by σ. If a value exactly equals to the mean of all the values of the feature, it will be normalized to 0. On the other hand, if it is below the mean, it will be a negative number, and positive number if above the mean. The size of those negative and positive numbers is determined by the standard deviation of the original feature. If the unnormalized data has a large standard deviation, the normalized values will be closer to 0.

After normalizing the data, the header rows is as follows:

	c_ax	c_ay	c_az	c_ecg	c_emg	c_eda	c_temp	c_resp
0	0.022271	-0.002078	-0.024127	-0.001559	-0.000277	0.091867	0.994664	-0.033454
1	0.022159	-0.002206	-0.023833	-0.006118	-0.000503	0.091576	0.993039	0.066222
2	0.022120	-0.002083	-0.023932	0.001807	0.000334	0.092113	0.993525	-0.057907
3	0.021939	-0.002362	-0.024162	-0.004786	0.000000	0.091921	0.991491	0.086033
4	0.022015	-0.002156	-0.024013	0.001578	-0.000233	0.091208	0.992649	0.072525
5	0.022152	-0.002335	-0.024065	0.000449	0.000011	0.087174	0.995091	-0.033463
6	0.022305	-0.002287	-0.023885	0.000509	-0.000208	0.086793	0.995215	-0.030672
7	0.022267	-0.002265	-0.024102	0.001219	-0.000154	0.087069	0.994542	-0.047134
8	0.022188	-0.002281	-0.023963	0.001176	-0.000186	0.086718	0.995664	0.007721
9	0.022211	-0.002126	-0.024261	-0.007183	-0.000081	0.086327	0.994822	-0.041691

PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is a common method for analyzing huge datasets with a high number of dimensions/features per observations, improving data interpretation while retaining the most information, and enabling the presentation of multidimensional data. Formally, PCA is a statistical method for lowering a dataset's dimensionality. To do this, the data are transformed linearly into a new coordinate system, where (most) of the variance in the data can be expressed with fewer dimensions than the initial data.

After applying Principal Component Analysis to the dataset, the variance of the predictor variables is obtained as follows:

```
Variance for PCA-1: 0.944268771218816
Variance for PCA-2: 0.03862736459798506
Variance for PCA-3: 0.013409294169775932
Variance for PCA-4: 0.0018904299362830878
```

DATA MINING MODELS / METHODS

LOGISTIC REGRESSION

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logistic regression model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula.

ADVANTAGES:

- The training of logistic regression is very effective and easier to implement and analyze. It is also very fast at classifying unknown records.
- Although it is less likely to do so, high-dimensional datasets can cause overfitting in logistic regression. To prevent over-fitting in these cases, one may want to consider regularization (L1 and L2) approaches.
- It can use model coefficients to determine the significance of a feature.

DISADVANTAGES:

- The assumption of linearity between the dependent variable and the independent variables is the main drawback of logistic regression.
- Logistic regression has a linear decision surface; hence it cannot address non-linear issues. Real-world situations rarely involve linearly separable data.

K-NEAREST NEIGHBOURS

The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another.

ADVANTAGES:

- KNN modeling is very time-efficient in terms of improvisation for a random modeling on the available data because it does not require a training period as the data itself is a model that will serve as the reference for future prediction.
- The only thing that needs to be calculated for KNN is the distance between various points using data from various features, and this distance can simply be calculated using distance formulas like Euclidian or Manhattan distances.

DISADVANTAGES:

- Poor performance with unbalanced data – If most of the data the model is trained on only contains one label, that label will be highly likely to be predicted.
- K value that is optimal — If K is selected wrong, the model will either be under- or overfit to the data.

RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

ADVANTAGES:

- It reduces overfitting in decision trees and helps to improve the accuracy
- It is flexible to both classification and regression problems
- It works well with both categorical and continuous values
- It automates missing values present in the data
- Normalising of data is not required as it uses a rule-based approach.

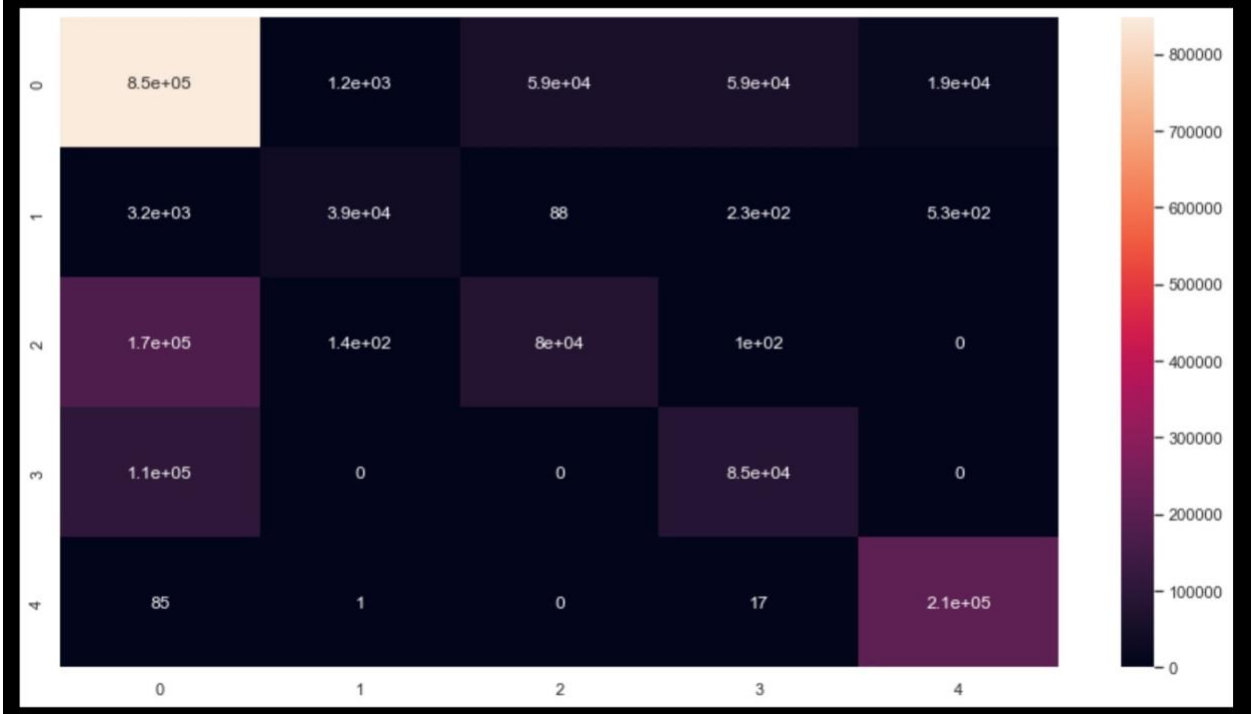
DISADVANTAGES:

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
- It also requires much time for training as it combines a lot of decision trees to determine the class.
- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

MODEL PERFORMANCE EVALUATION & VISUALIZATIONS

CONFUSION MATRIX:

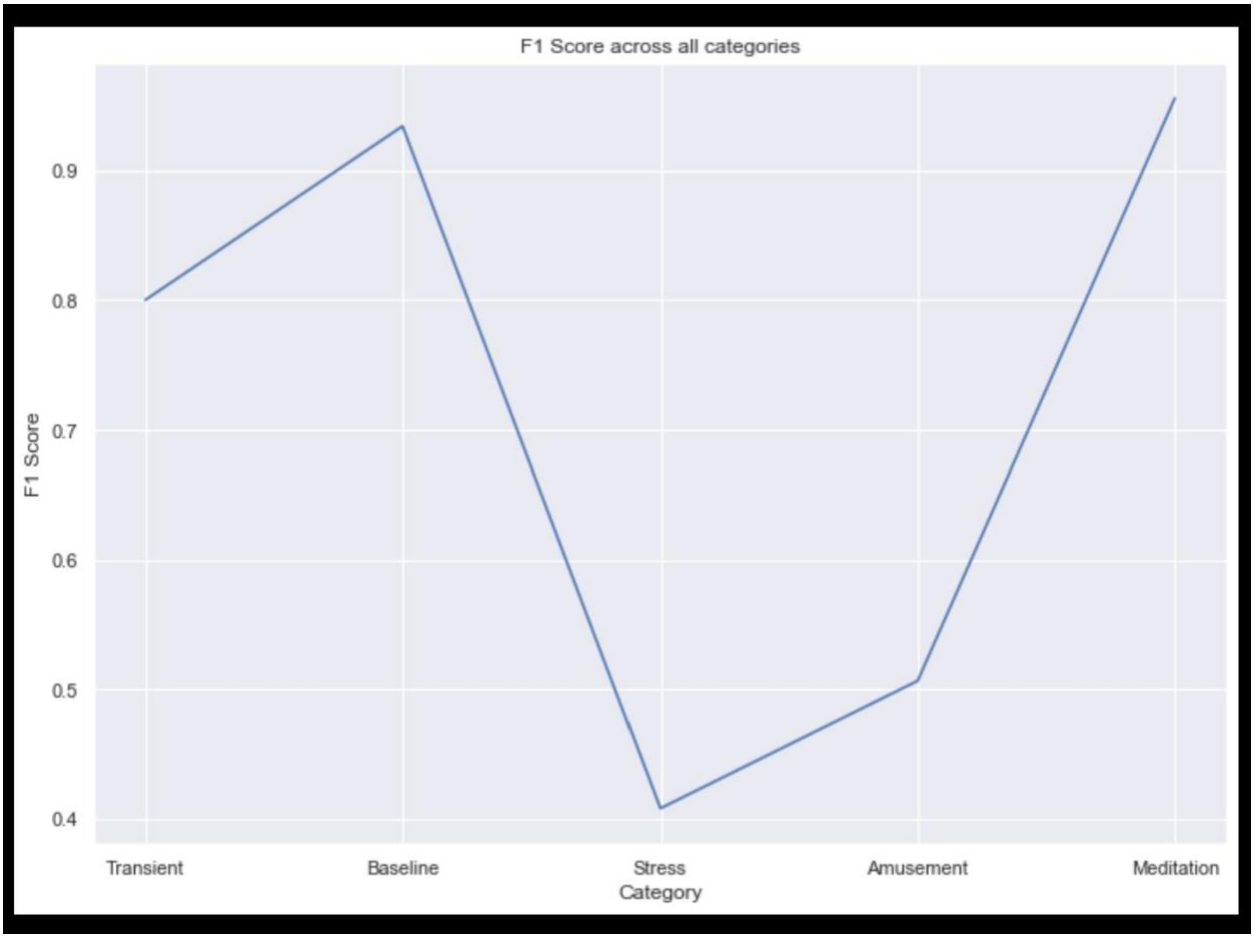
```
Confusion Matrix of training set
[[847202   1212  59205  58538  18704]
 [  3200  38807     88    229    528]
 [173610    143  79578    102      0]
 [105740      0      0  85369      0]
 [    85      1      0     17 208006]]
```



CLASSIFICATION SUMMARY REPORT:

Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.7485	0.8596	0.8002	245572
1.0	0.9641	0.9065	0.9344	10807
2.0	0.5751	0.3163	0.4081	63502
3.0	0.5909	0.4434	0.5066	48082
4.0	0.9159	0.9996	0.9559	52129
accuracy			0.7484	420092
macro avg	0.7589	0.7051	0.7211	420092
weighted avg	0.7306	0.7484	0.7301	420092



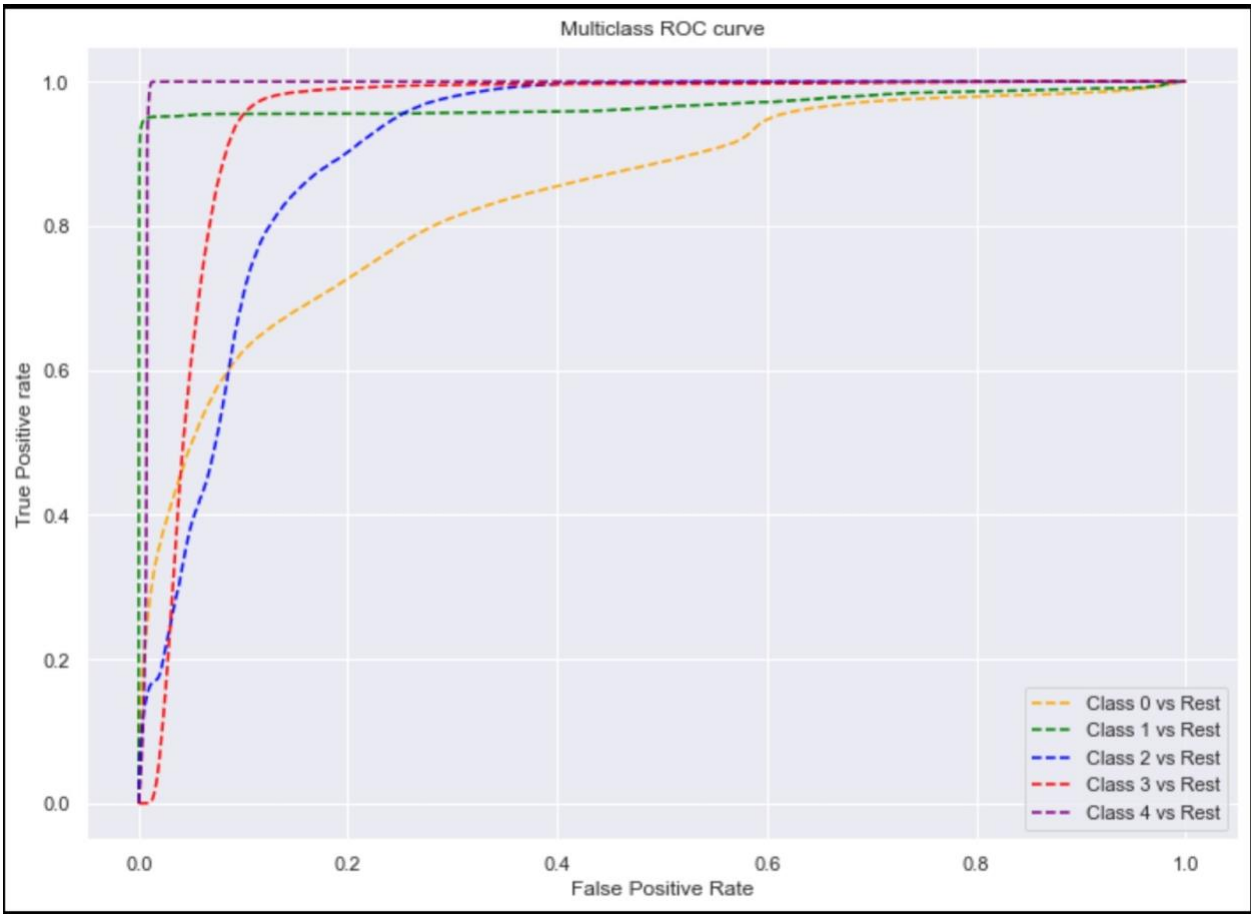
ROC Curve:

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

False Positive Rate (**FPR**) is defined as follows:



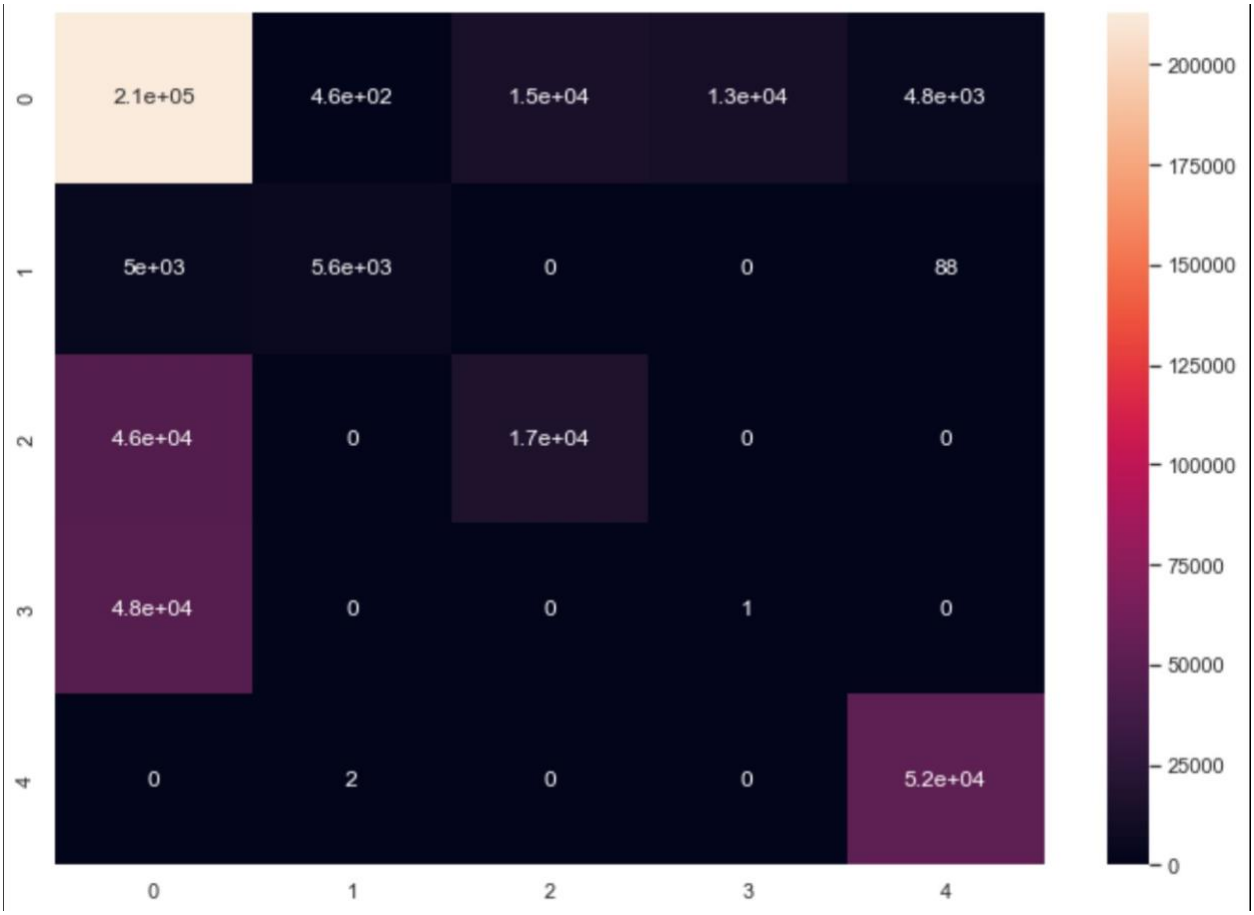
Interpretations:

From the above we observe that:

- The accuracy is 74% for the Logistic Regression Model implemented
- The F1 score for ‘Stress’ category is the lowest.
- The categories ‘Baseline’ and ‘Meditation’ have the highest F1 Score
- The ROC Curve for the ‘Transient’ category is not ideal, whereas for ‘Meditation’ & ‘Baseline’ we obtained the perfect ROC Curve.

LOGISTIC REGRESSION ON PCA WITH 4 PRINCIPAL COMPONENTS

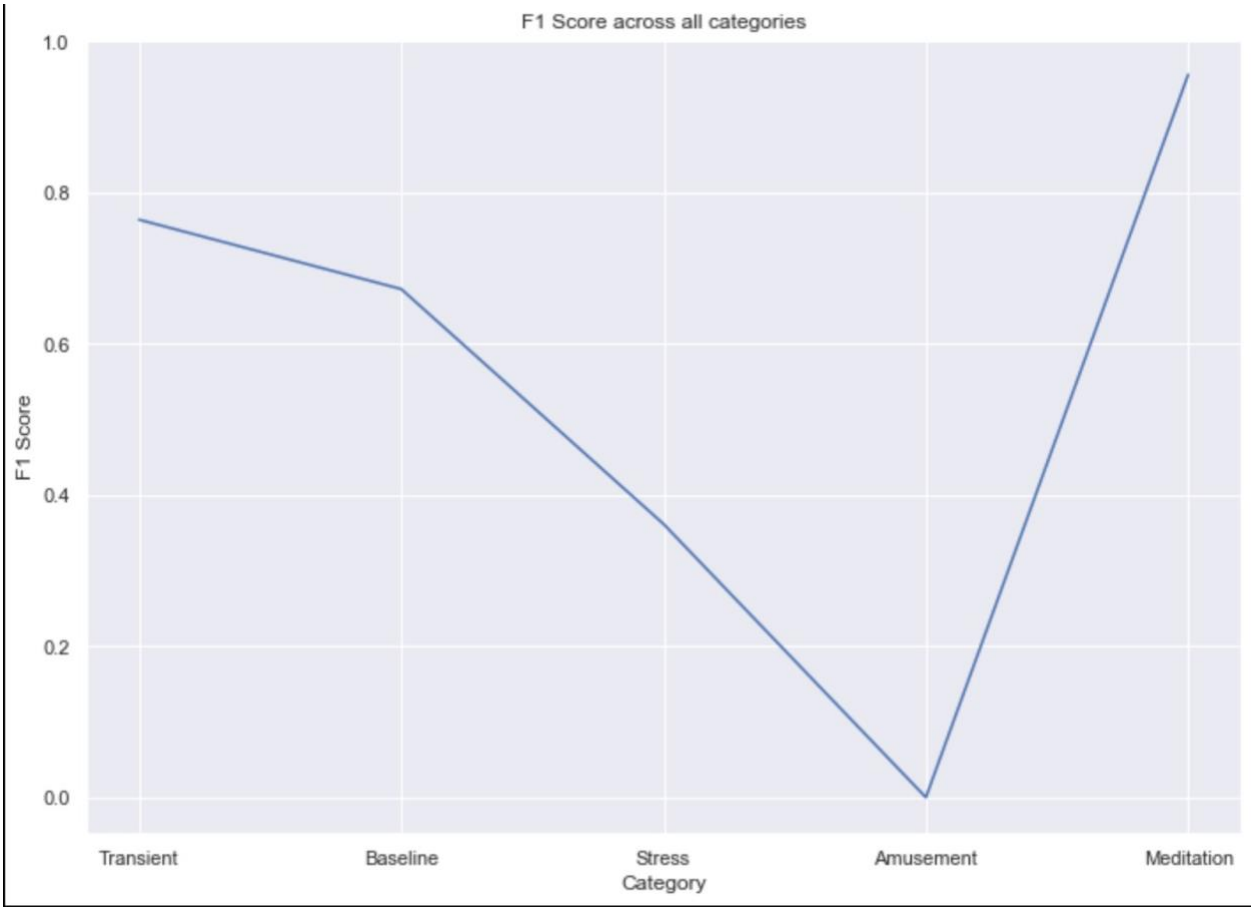
After application of Logistic Regression on PCA Transformed data with 4 Principal Components, the following is the Confusion Matrix:



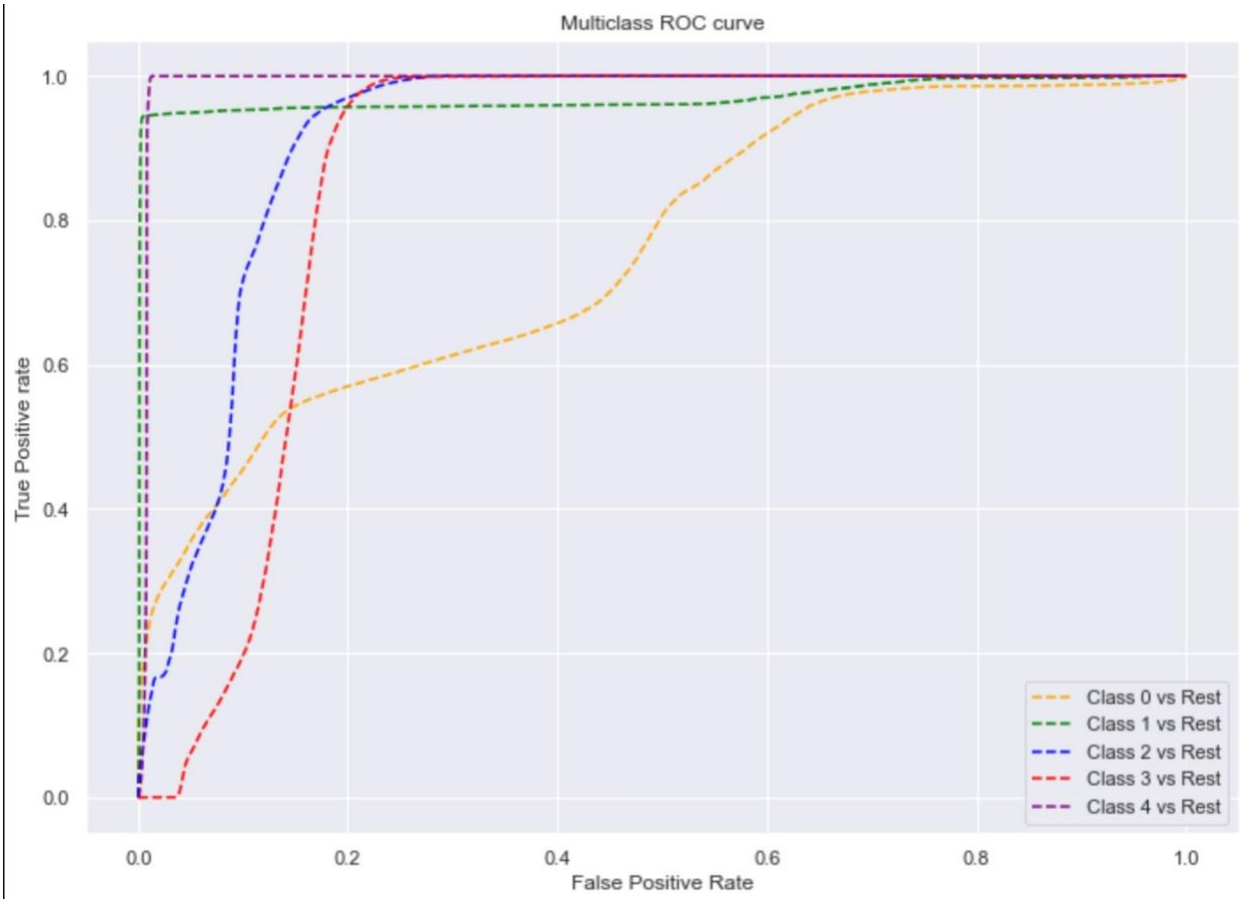
CLASSIFICATION SUMMARY REPORT:

Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.6823	0.8673	0.7637	245683
1.0	0.9249	0.5277	0.6720	10694
2.0	0.5413	0.2712	0.3613	63532
3.0	0.0001	0.0000	0.0000	47951
4.0	0.9148	1.0000	0.9555	52232
accuracy			0.6860	420092
macro avg	0.6127	0.5332	0.5505	420092
weighted avg	0.6182	0.6860	0.6372	420092



ROC Curve:

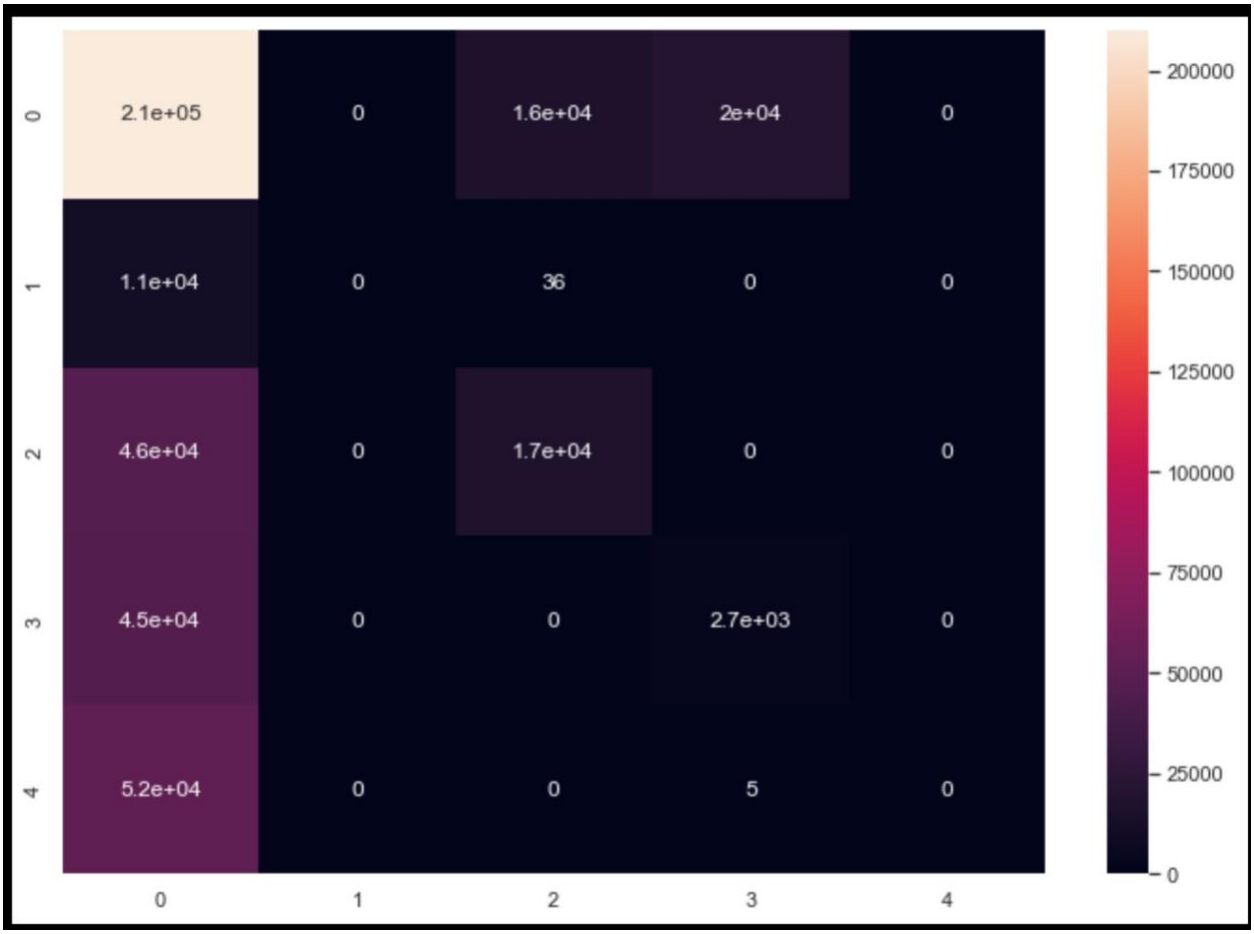


Interpretations:

From the above we observe that:

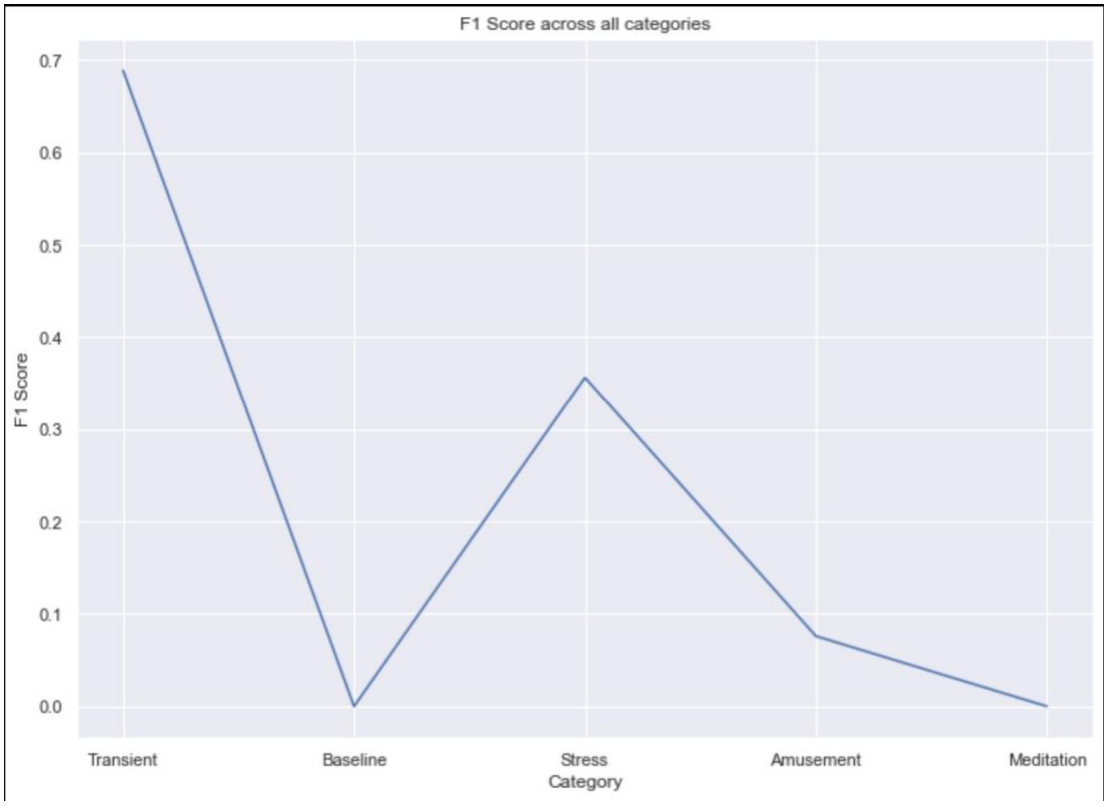
- The accuracy is 68.60% for the above implemented model
- The F1 score for ‘Amusement’ category is the lowest.
- The category ‘Meditation’ continues to have the highest F1 Score in this model
- The ROC Curve for the ‘Transient’ category is not ideal, whereas for ‘Meditation’ & ‘Baseline’ we obtained the perfect ROC Curve.

LOGISTIC REGRESSION WITH PCA WITH TWO PRINCIPAL COMPONENTS

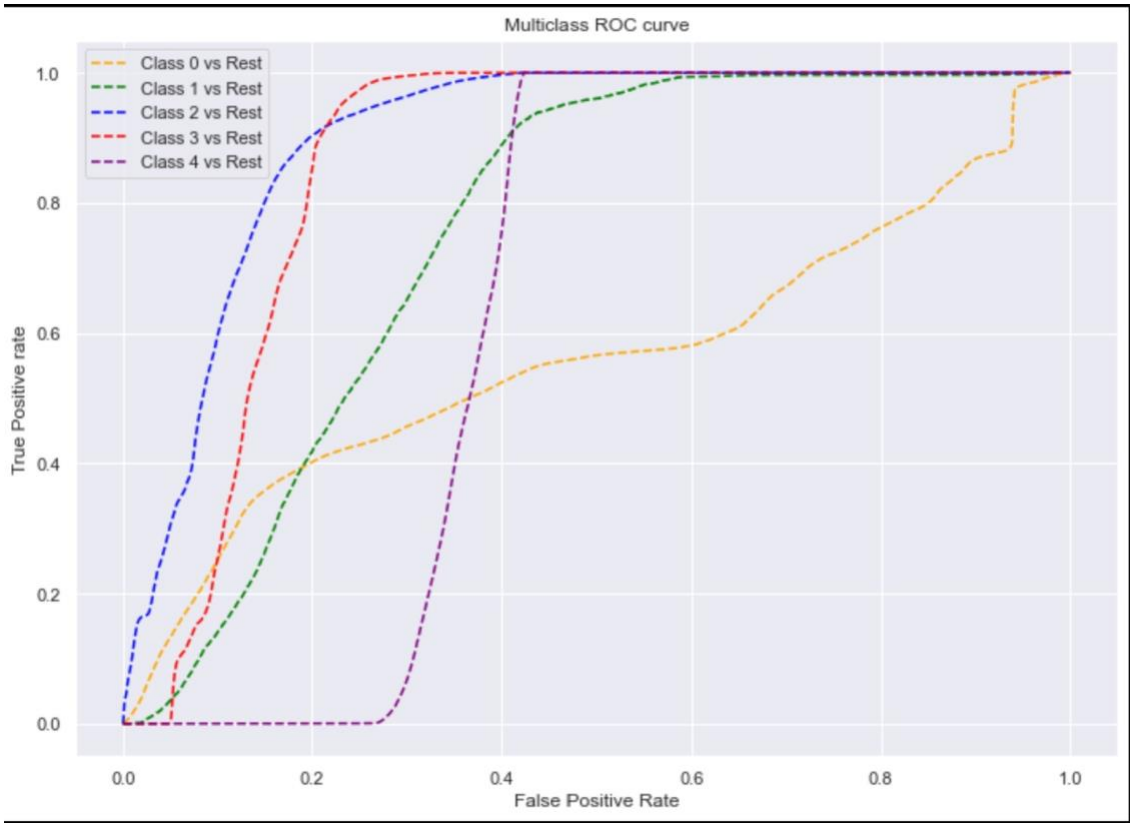


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.5763	0.8556	0.6887	245683
1.0	0.0000	0.0000	0.0000	10694
2.0	0.5224	0.2704	0.3563	63532
3.0	0.1190	0.0558	0.0760	47951
4.0	0.0000	0.0000	0.0000	52232
accuracy			0.5476	420092
macro avg	0.2435	0.2364	0.2242	420092
weighted avg	0.4296	0.5476	0.4653	420092


F1 Score across all categories
F1 Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Transient
Baseline
Stress
Amusement
Meditation
Category

ROC CURVE:

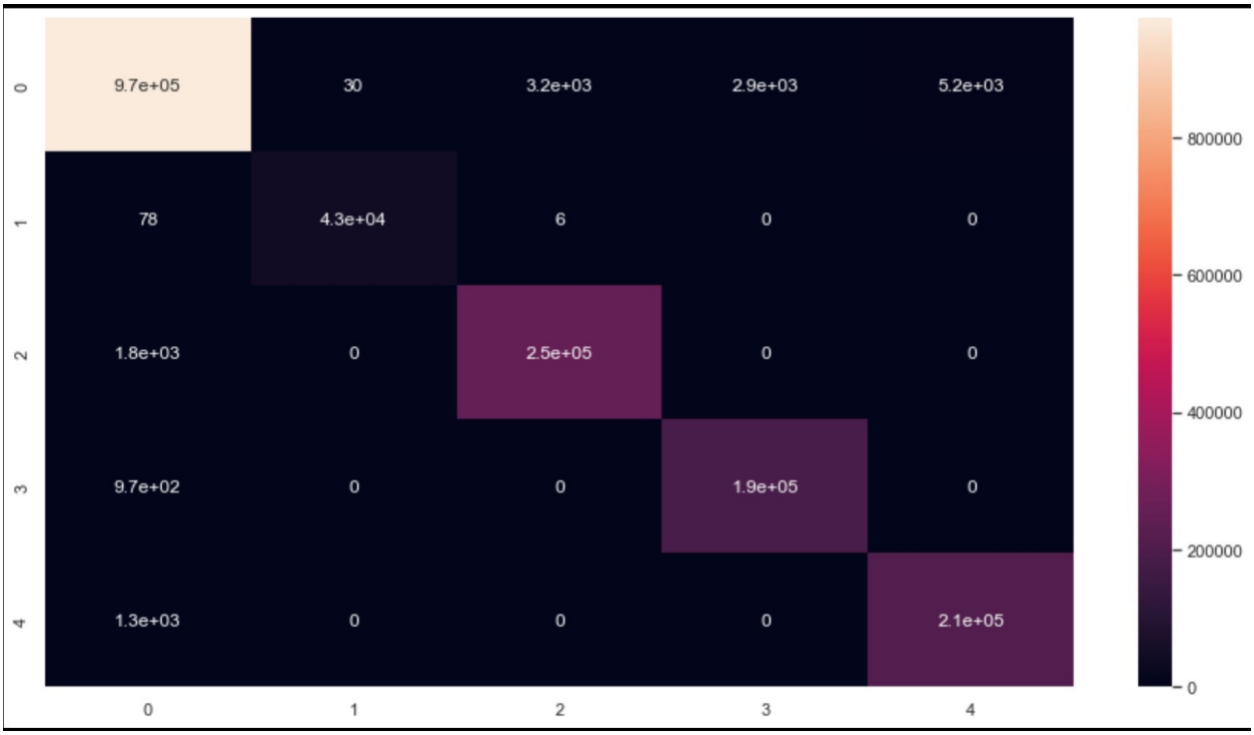

Multiclass ROC curve
Class 0 vs Rest
Class 1 vs Rest
Class 2 vs Rest
Class 3 vs Rest
Class 4 vs Rest
True Positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

Interpretations:

From the above we observe that:

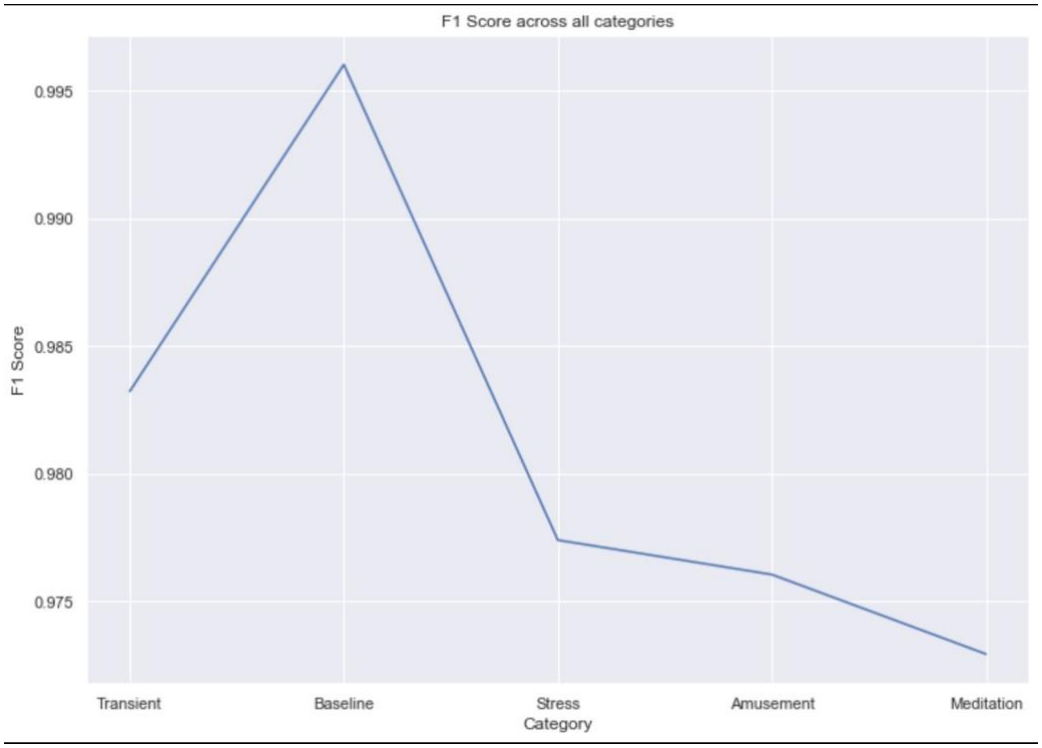
- The accuracy falls to 54.76% for the above implemented model
- The F1 score for ‘Baseline’ and ‘Meditation’ categories is the lowest.
- The category ‘Transient’ has the highest F1 Score in this model
- The ROC Curve yields bad results on all the categories present, and is not ideal in either of them.

KNN Model for K = 3

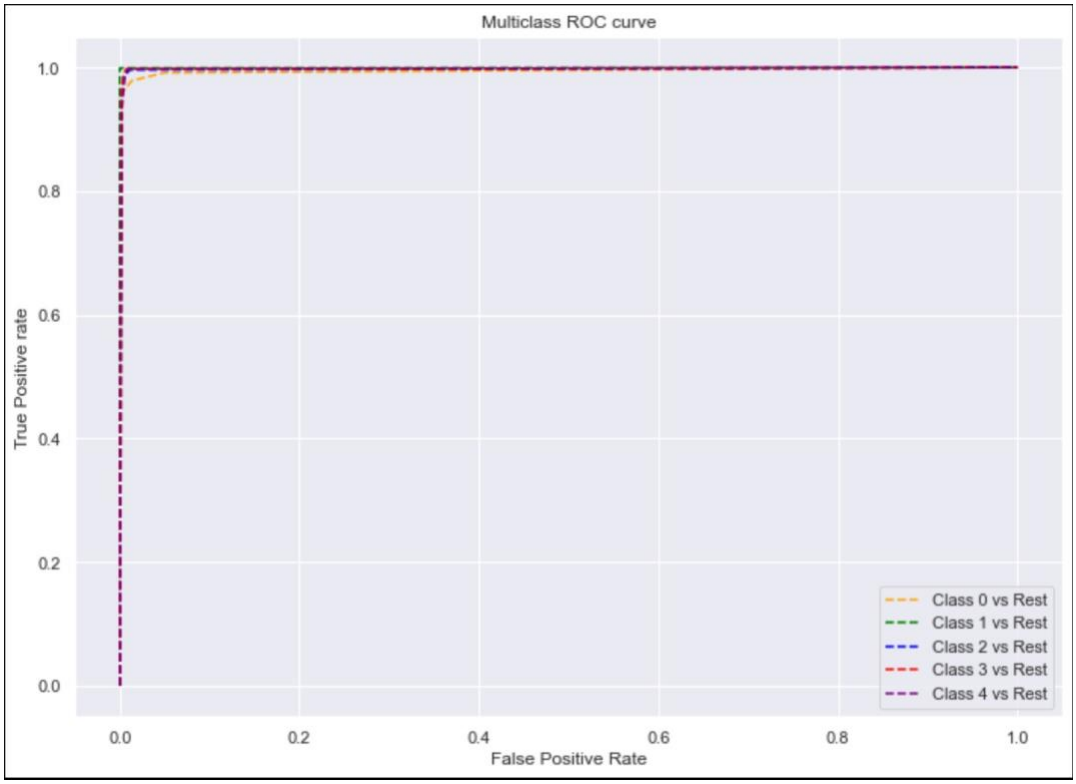


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9904	0.9786	0.9845	245572
1.0	0.9986	0.9966	0.9976	10807
2.0	0.9747	0.9853	0.9800	63502
3.0	0.9691	0.9877	0.9783	48082
4.0	0.9606	0.9853	0.9728	52129
accuracy			0.9819	420092
macro avg	0.9787	0.9867	0.9826	420092
weighted avg	0.9821	0.9819	0.9820	420092



ROC Curve:

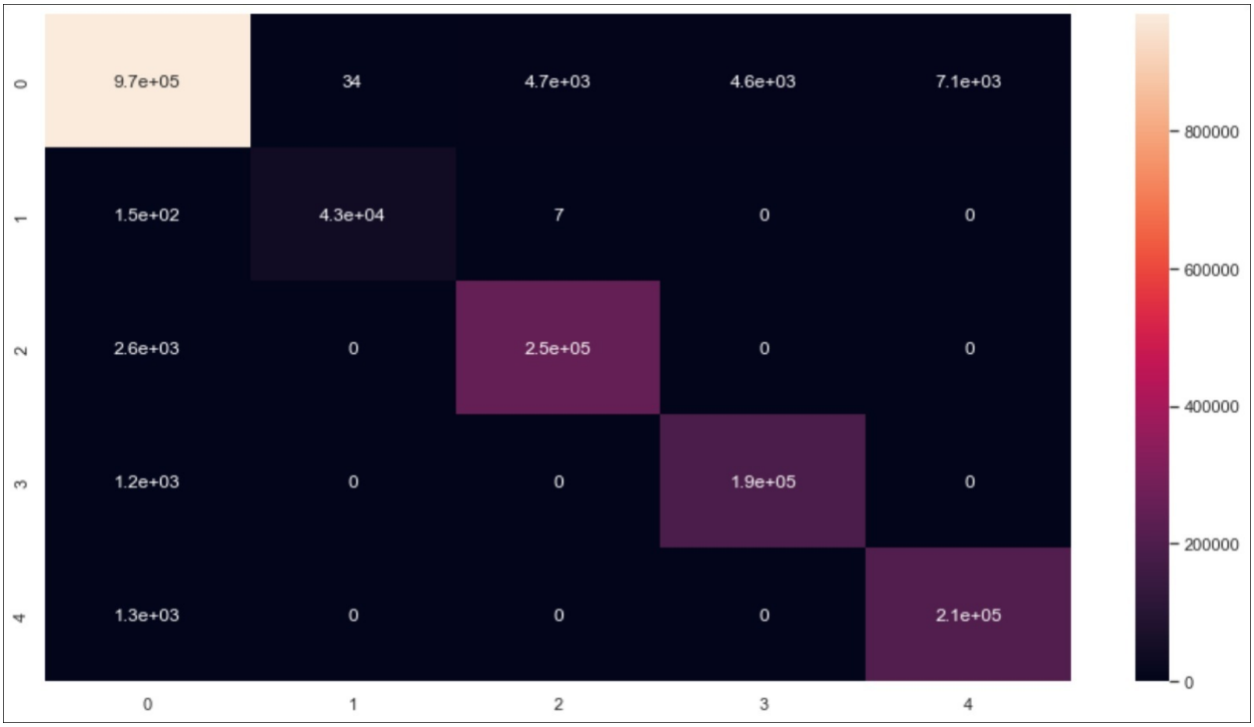


Interpretations:

From the above we observe that:

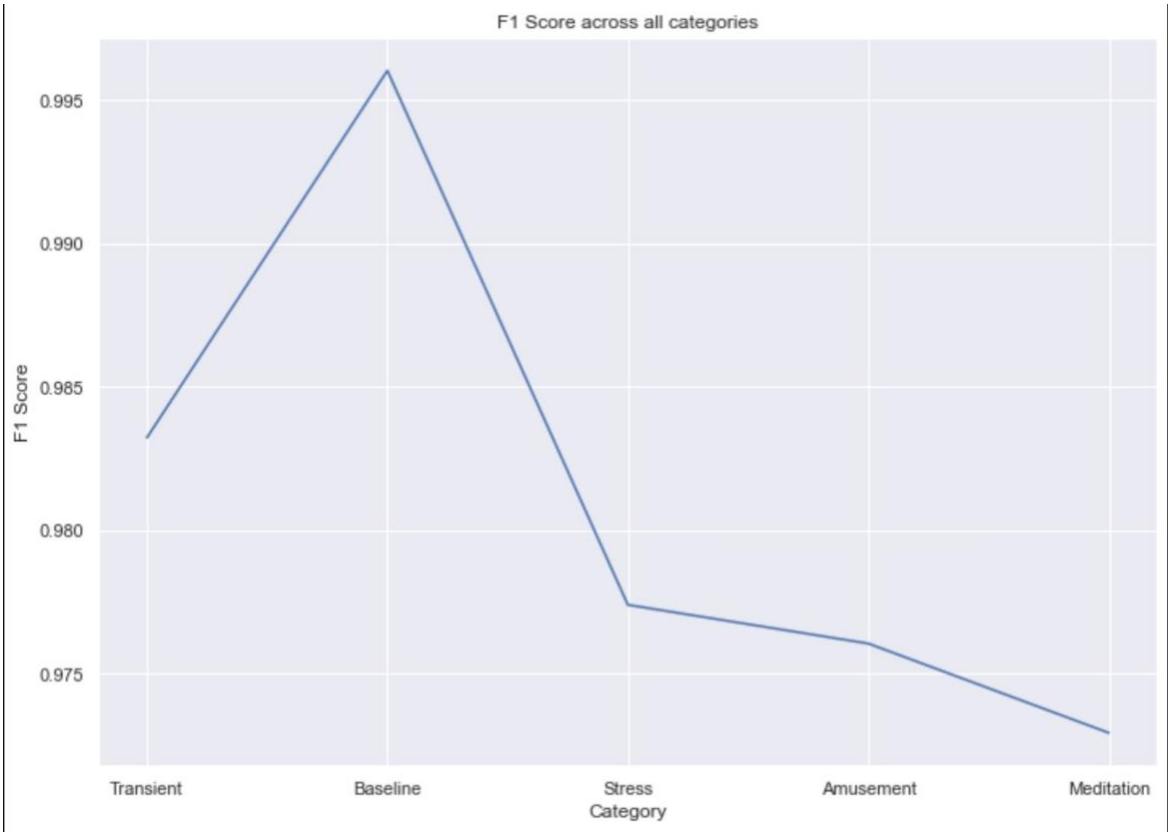
- The KNN Model with K=3, yields good results, with an accuracy of 98.19%
- The F1 score for ‘Meditation’ category is the lowest.
- The category ‘Baseline’ has the highest F1 Score in this model
- The ROC Curve continues to yield great results on all the categories present

KNN for K = 5

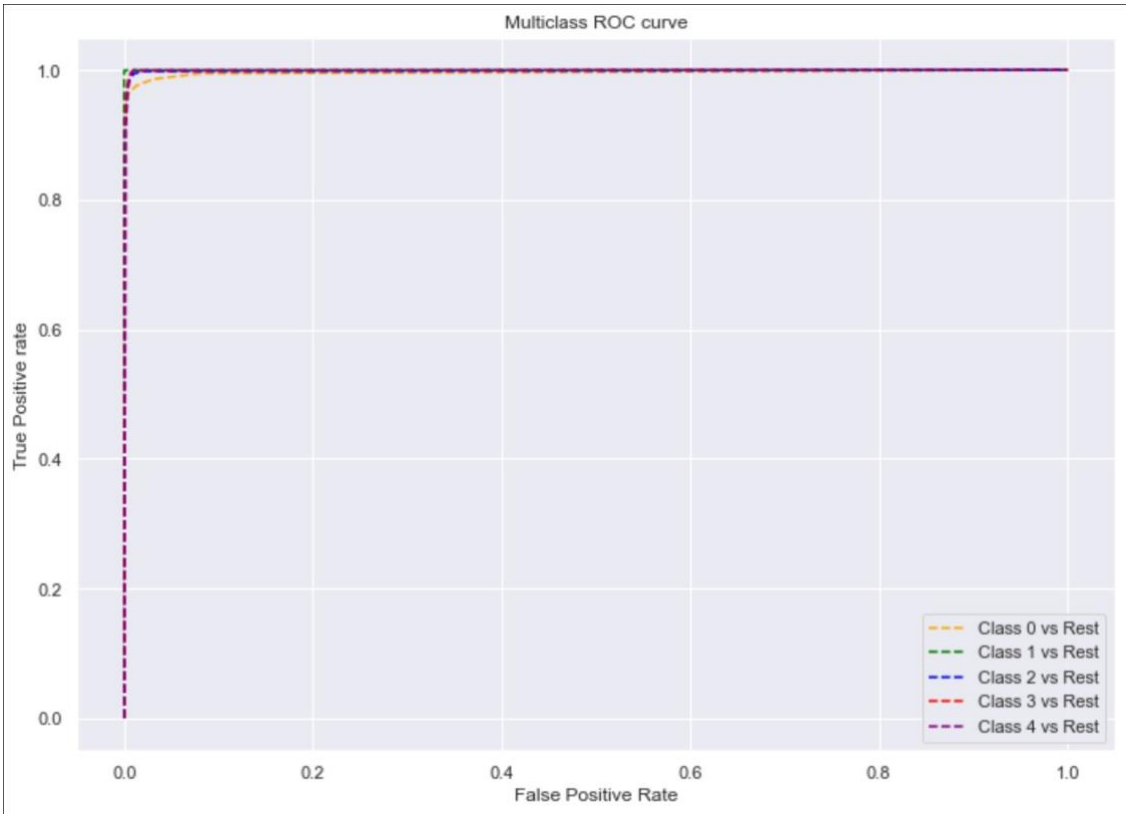


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9910	0.9756	0.9832	245572
1.0	0.9982	0.9938	0.9960	10807
2.0	0.9705	0.9844	0.9774	63502
3.0	0.9638	0.9887	0.9761	48082
4.0	0.9574	0.9890	0.9729	52129
accuracy			0.9805	420092
macro avg	0.9762	0.9863	0.9811	420092
weighted avg	0.9808	0.9805	0.9806	420092


F1 Score across all categories
F1 Score
0.995
0.990
0.985
0.980
0.975
Transient
Baseline
Stress
Amusement
Meditation
Category

ROC Curve:

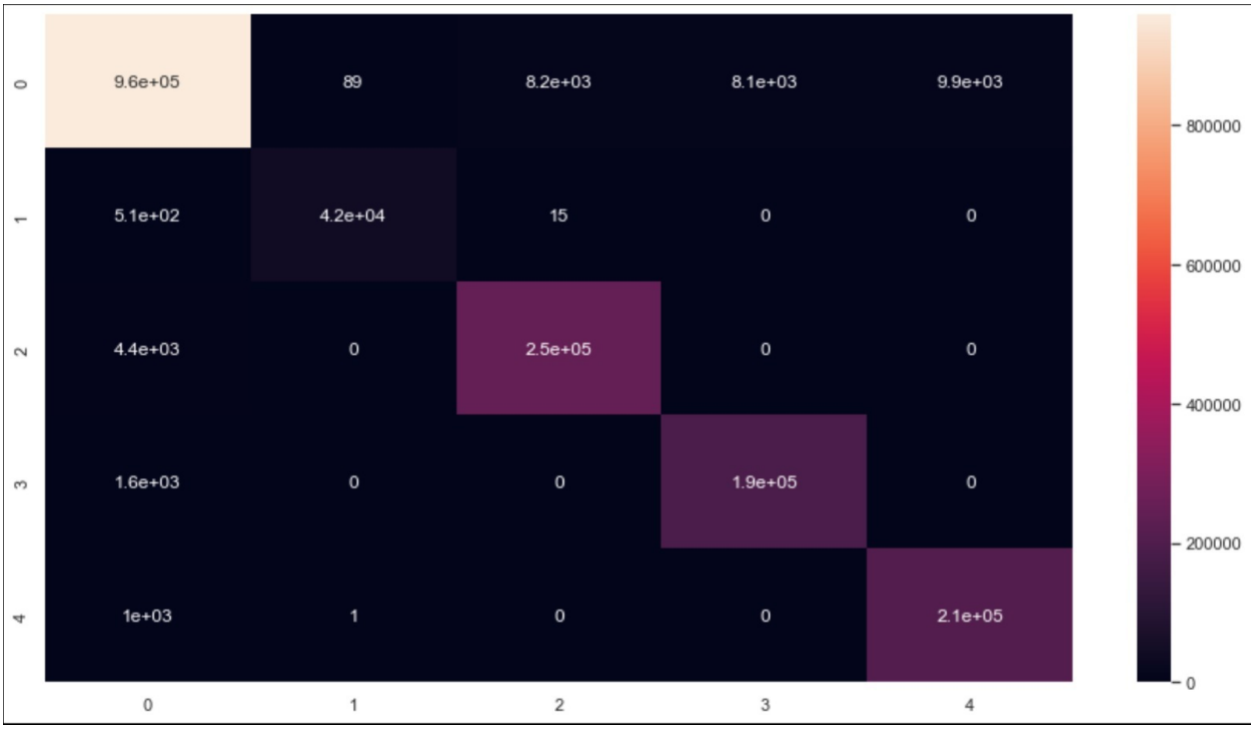

Multiclass ROC curve
True Positive rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Class 0 vs Rest
Class 1 vs Rest
Class 2 vs Rest
Class 3 vs Rest
Class 4 vs Rest

Interpretations:

From the above we observe that:

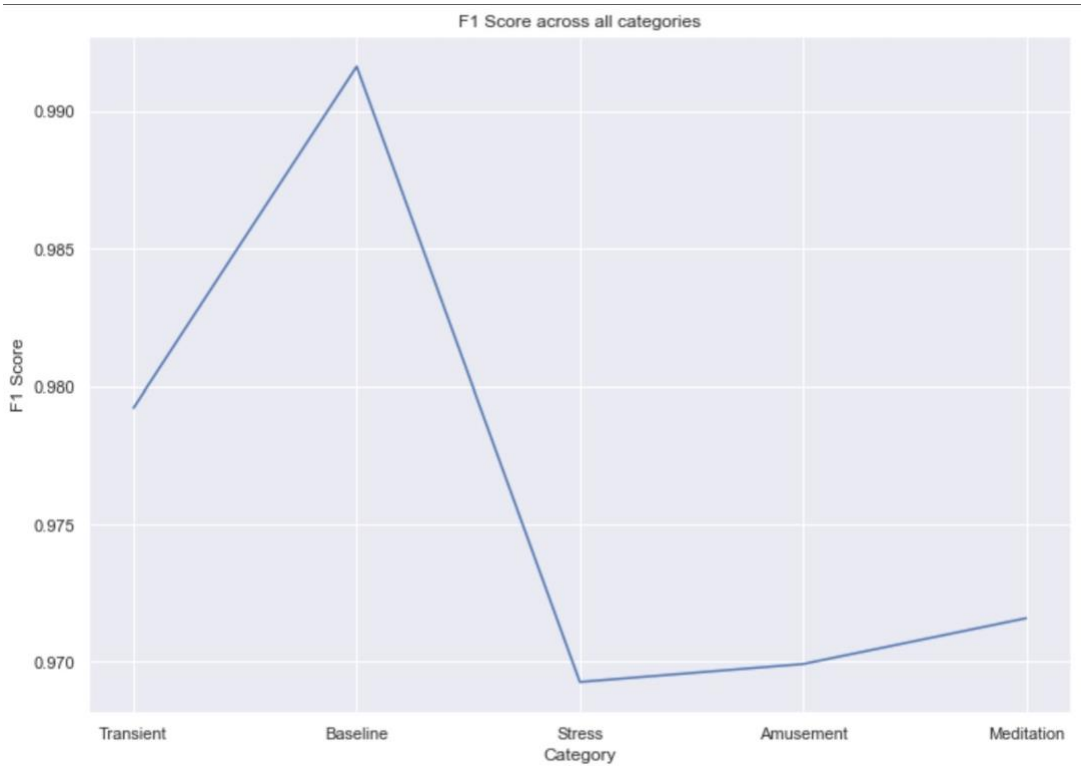
- The KNN Model with K=5, yields good results, with an accuracy of 98.05%
- The F1 scores for this model are similar to that of K=3
- The ROC Curve continues to generate ideal results.

KNN for K = 11

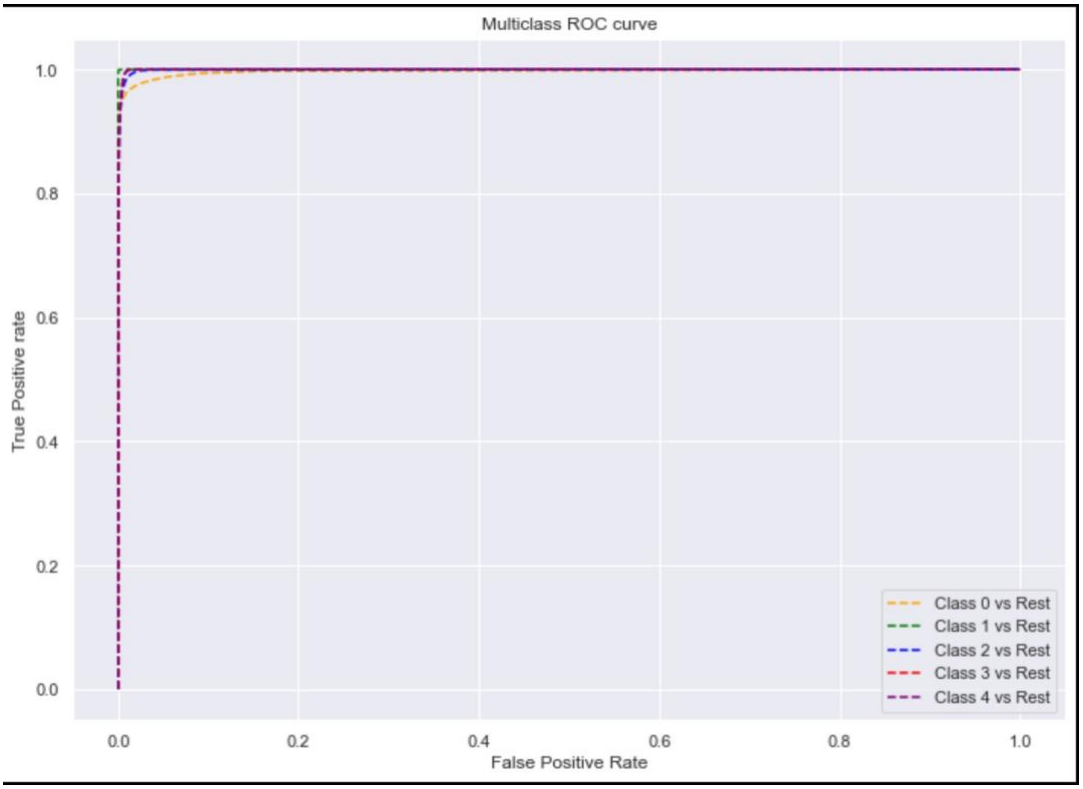


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9901	0.9686	0.9792	245572
1.0	0.9969	0.9864	0.9916	10807
2.0	0.9603	0.9784	0.9693	63502
3.0	0.9515	0.9891	0.9699	48082
4.0	0.9507	0.9934	0.9716	52129
accuracy			0.9760	420092
macro avg	0.9699	0.9832	0.9763	420092
weighted avg	0.9764	0.9760	0.9760	420092



ROC Curve:

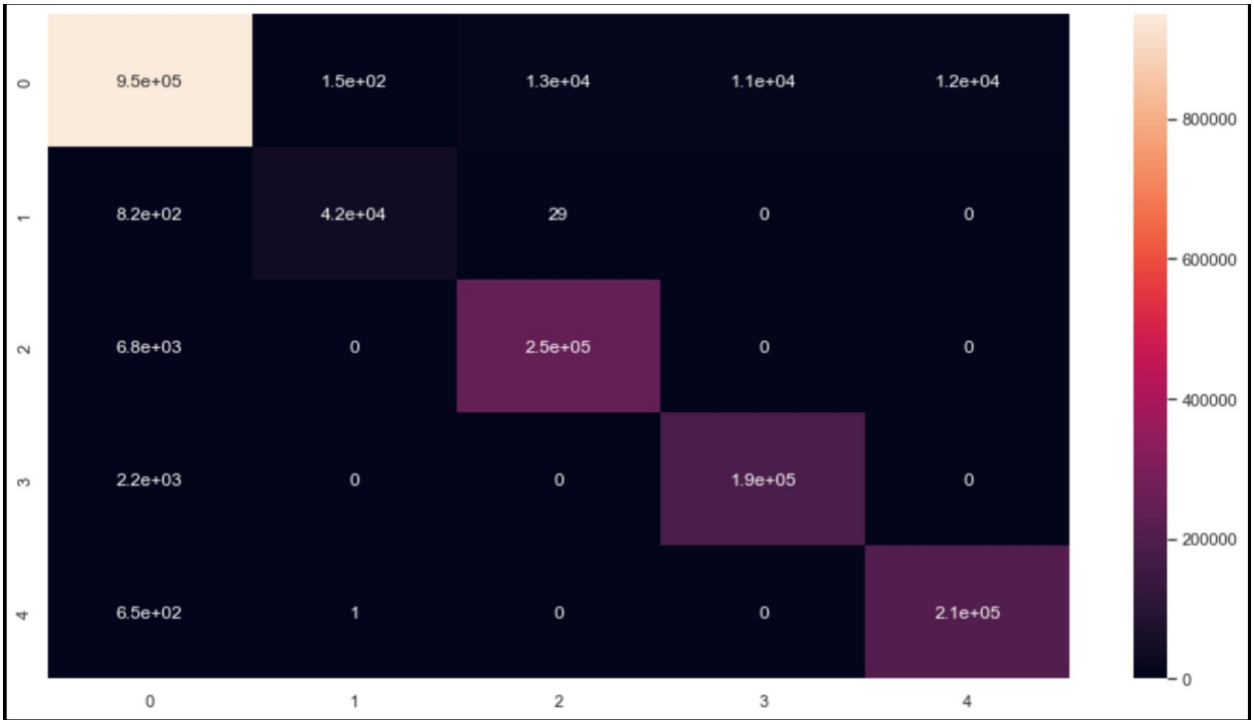


Interpretations:

From the above we observe that:

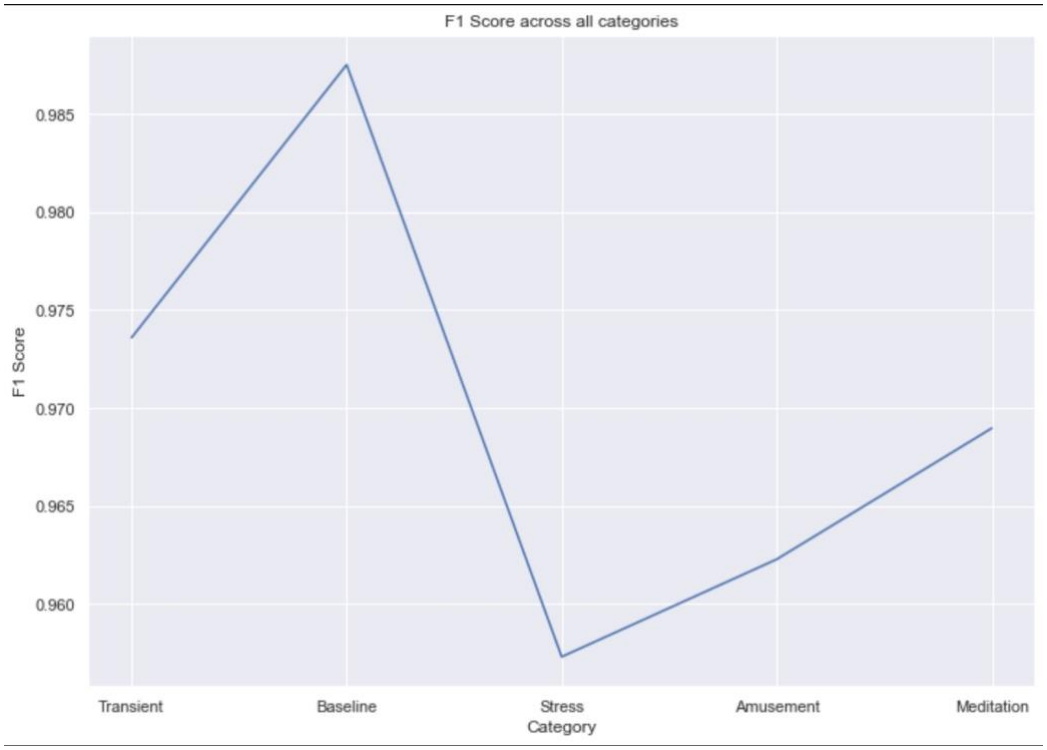
- The KNN Model with K=11, yields good results, with an accuracy of 97.6%
- 'Stress' category has the lowest F1 score, while 'Baseline' category has the highest.
- The ROC Curve is similar to that of previous k values of all KNN models

KNN for K = 21

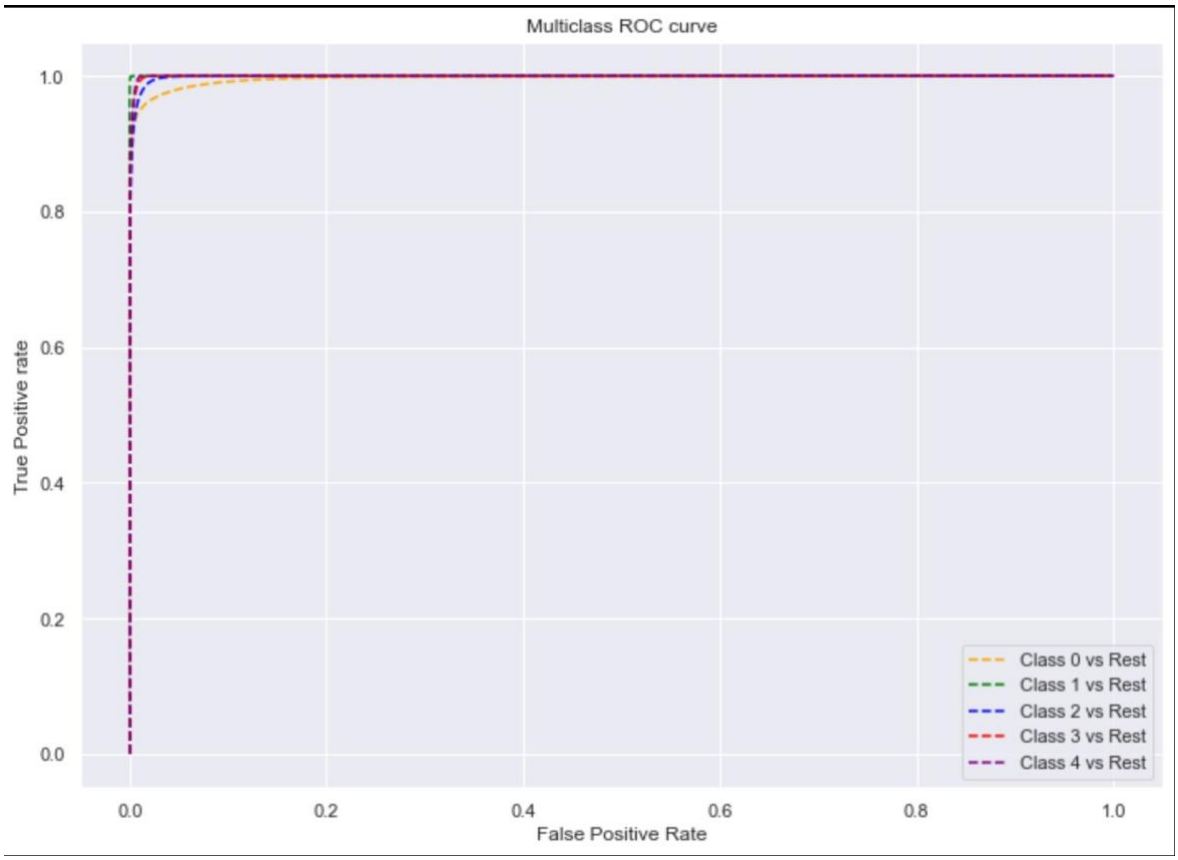


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9875	0.9600	0.9736	245572
1.0	0.9955	0.9796	0.9875	10807
2.0	0.9458	0.9691	0.9573	63502
3.0	0.9385	0.9874	0.9623	48082
4.0	0.9432	0.9962	0.9690	52129
accuracy			0.9695	420092
macro avg	0.9621	0.9785	0.9699	420092
weighted avg	0.9703	0.9695	0.9696	420092


F1 Score across all categories
F1 Score
0.985
0.980
0.975
0.970
0.965
0.960
Transient
Baseline
Stress
Amusement
Meditation
Category

ROC Curve:

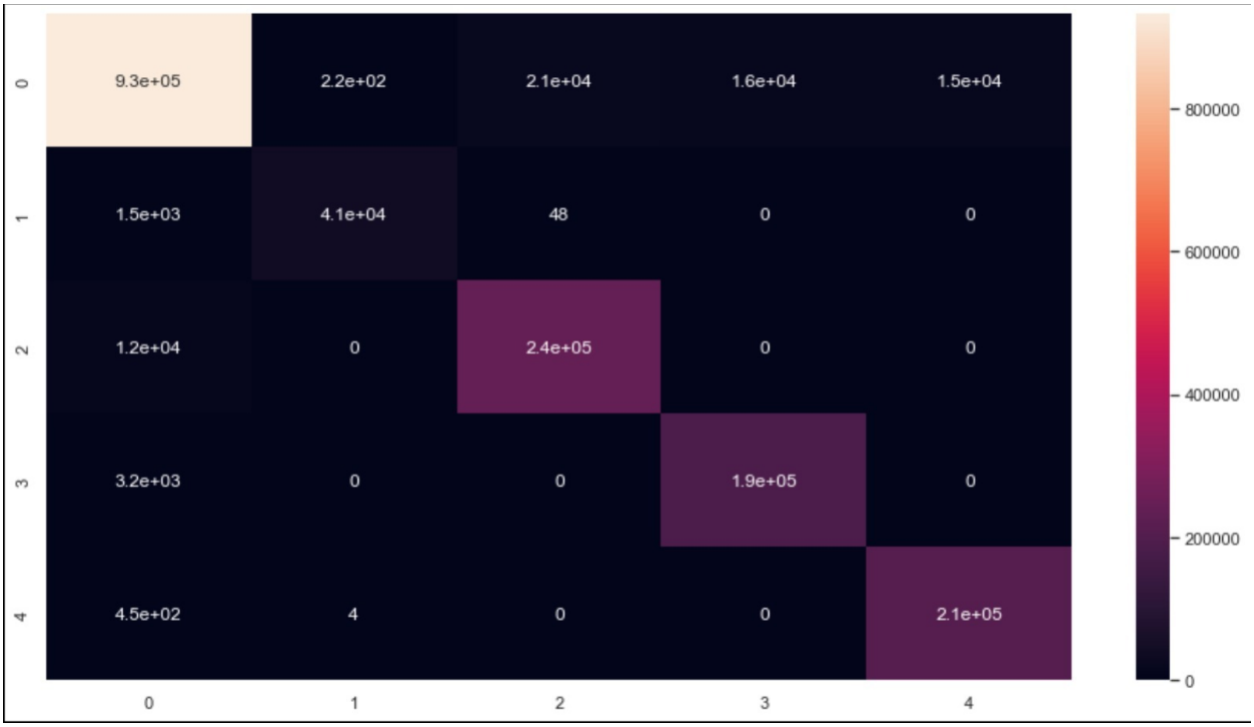

Multiclass ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 vs Rest
Class 1 vs Rest
Class 2 vs Rest
Class 3 vs Rest
Class 4 vs Rest

Interpretations:

From the above we observe that:

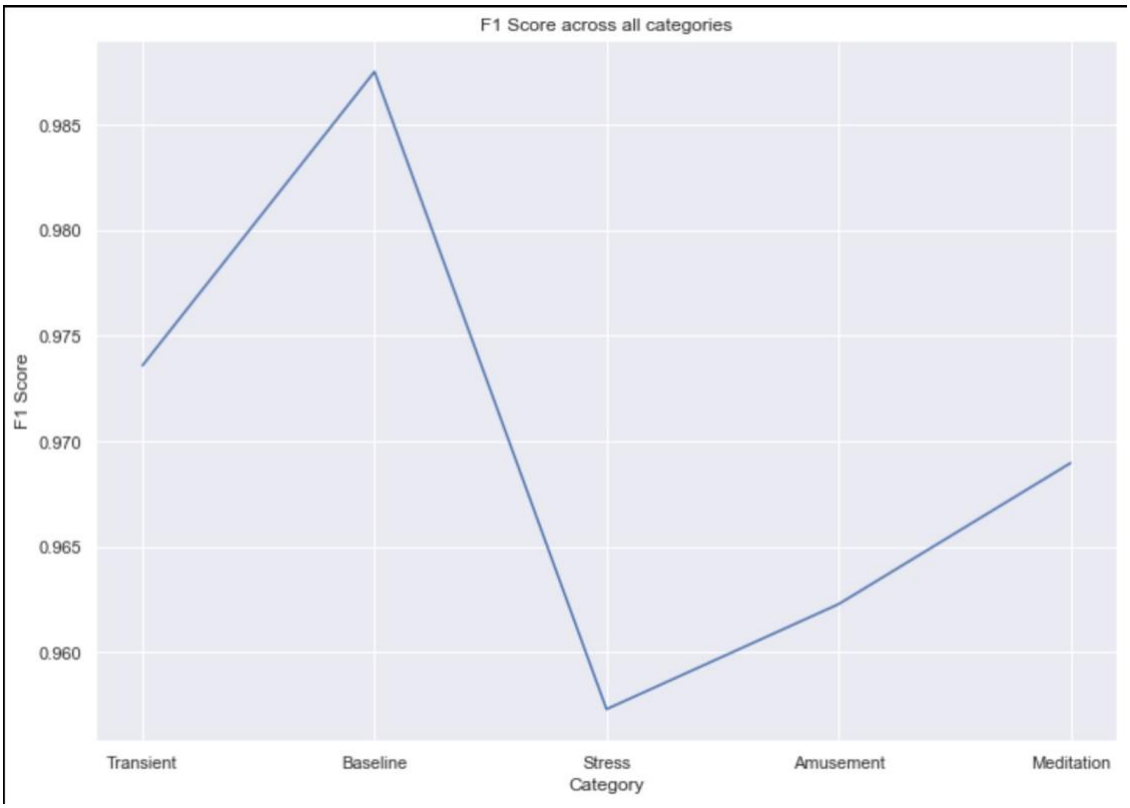
- The KNN Model with K=21 gives an accuracy of 96.95%
- The F1 score is lowest for ‘Stress’ and highest for ‘Baseline’ category
- The ROC Curve still holds ideal results on all the categories present

KNN for K = 51

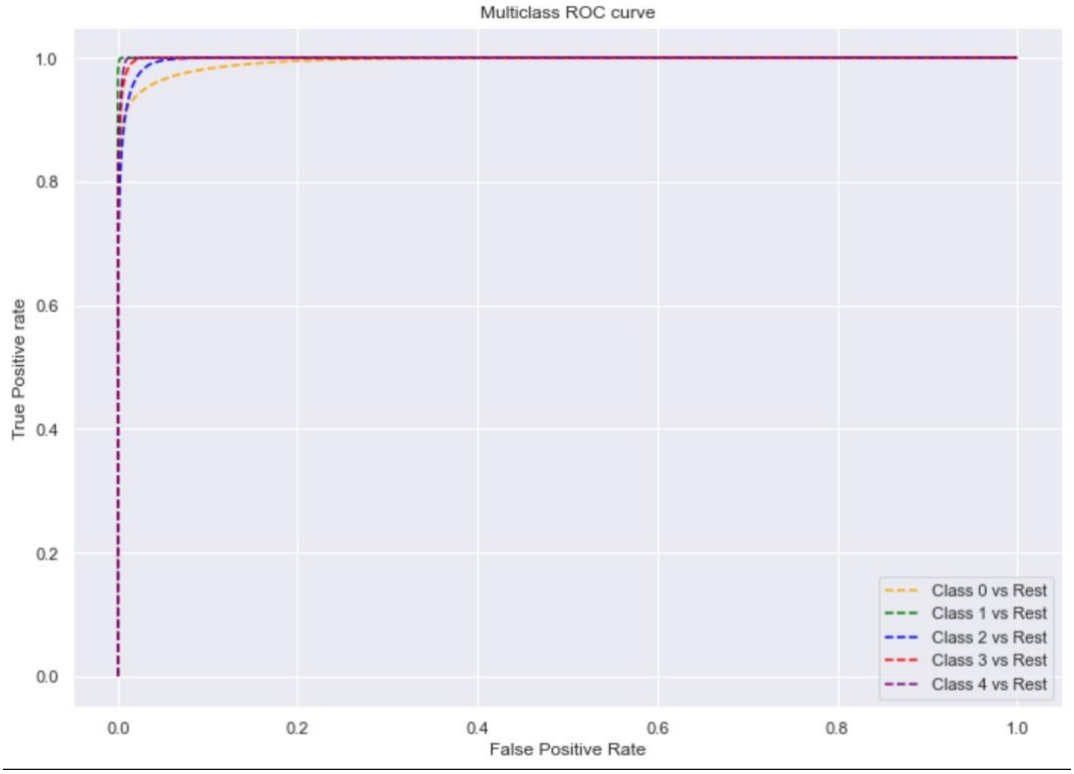


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9811	0.9448	0.9626	245572
1.0	0.9948	0.9619	0.9781	10807
2.0	0.9158	0.9508	0.9330	63502
3.0	0.9177	0.9826	0.9490	48082
4.0	0.9330	0.9980	0.9644	52129
accuracy			0.9571	420092
macro avg	0.9485	0.9676	0.9574	420092
weighted avg	0.9584	0.9571	0.9572	420092



ROC Curve:

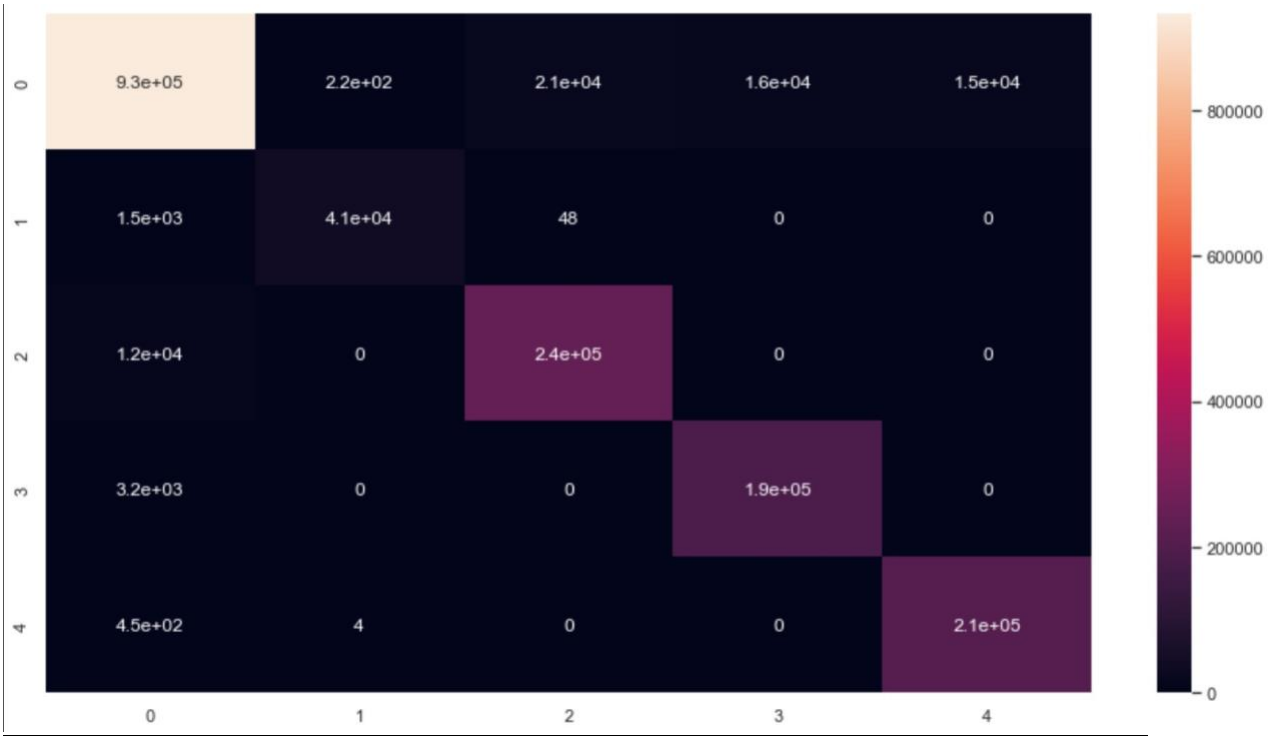


Interpretations:

From the above we observe that:

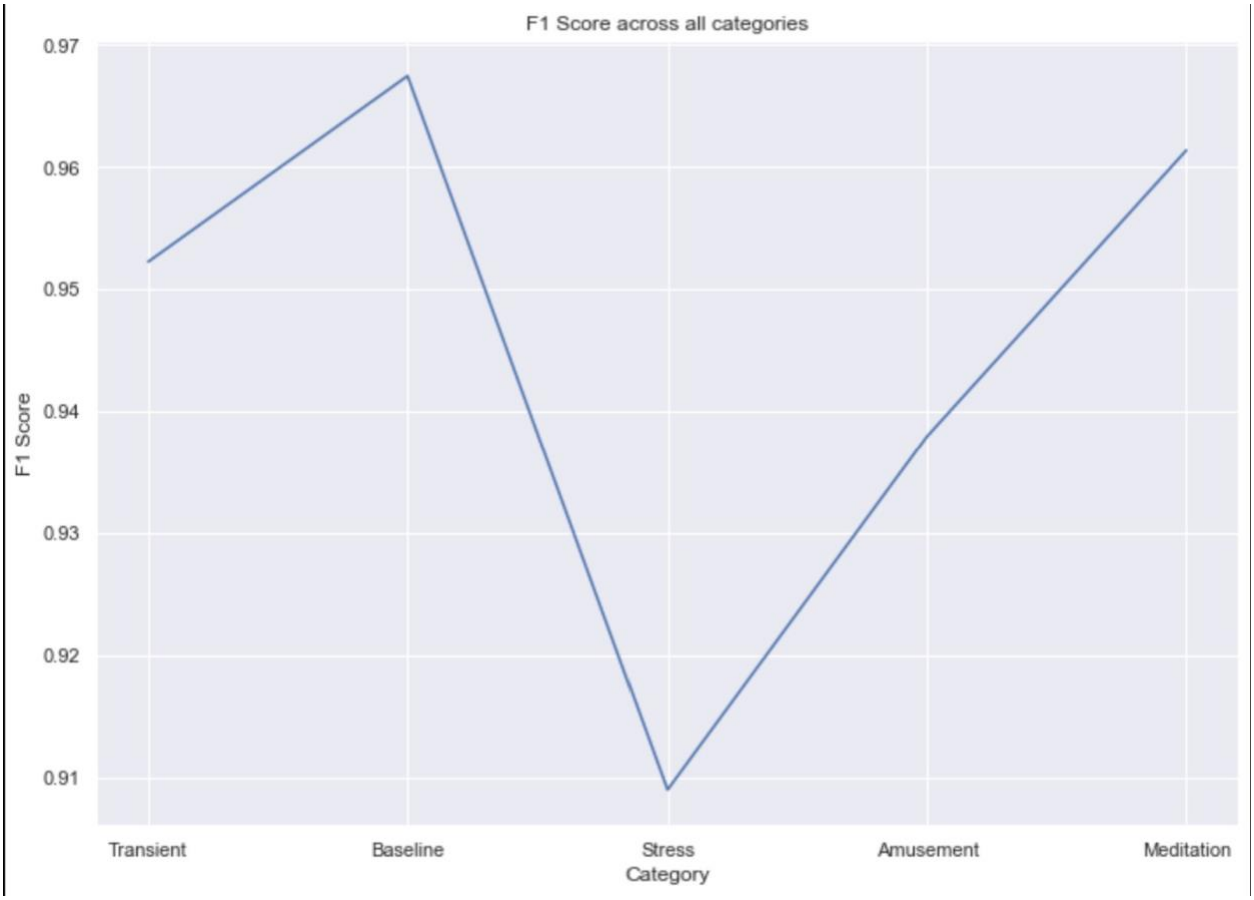
- The KNN Model with K=51 gives an accuracy of 95.71%
- The F1 score is lowest for ‘Stress’ and highest for ‘Baseline’ categories
- The ROC Curve looks ideal following the previous models.

KNN for K = 100

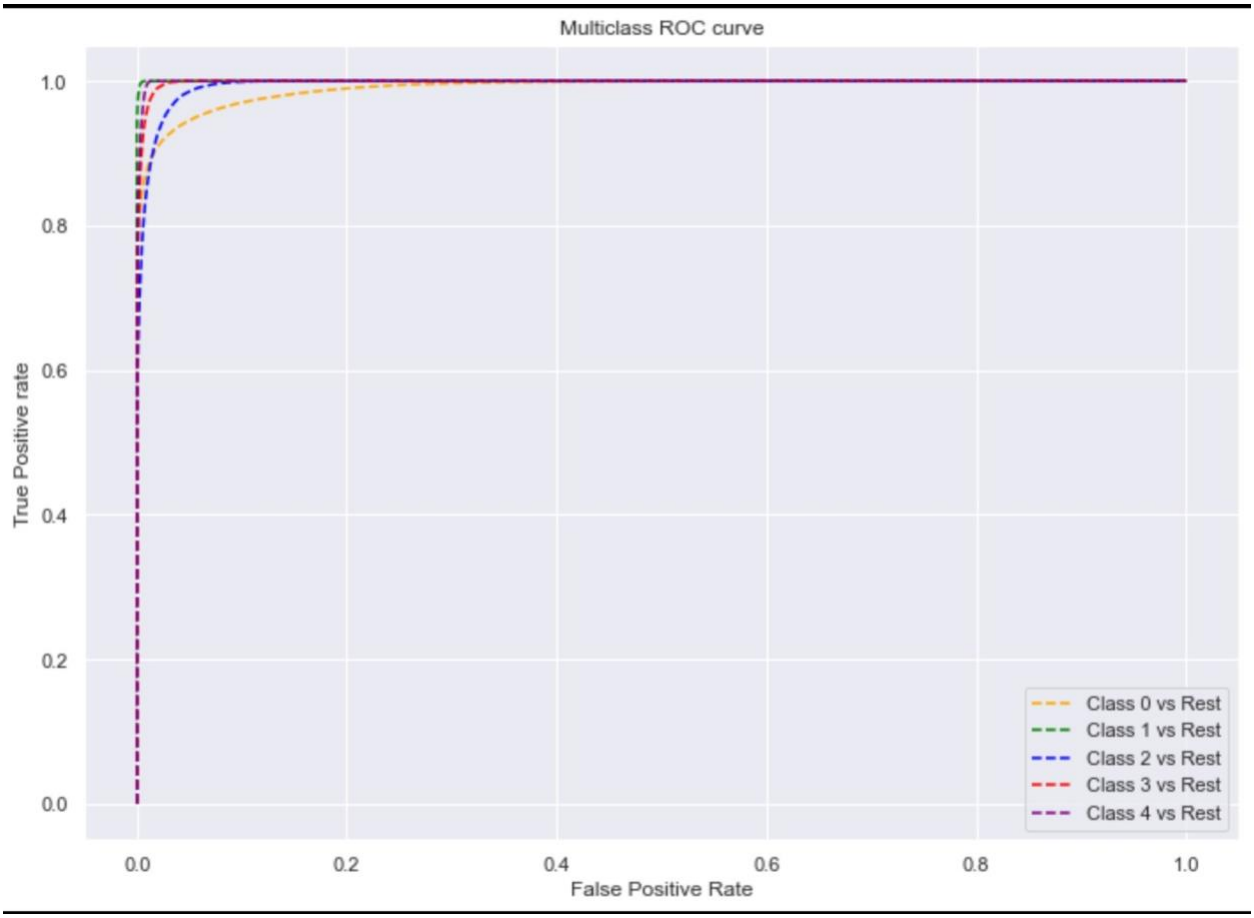


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9745	0.9309	0.9522	245572
1.0	0.9910	0.9449	0.9674	10807
2.0	0.8864	0.9327	0.9090	63502
3.0	0.9007	0.9783	0.9379	48082
4.0	0.9268	0.9985	0.9613	52129
accuracy			0.9454	420092
macro avg	0.9359	0.9571	0.9456	420092
weighted avg	0.9473	0.9454	0.9456	420092


F1 Score across all categories
0.97
0.96
0.95
0.94
0.93
0.92
0.91
F1 Score
Transient
Baseline
Stress
Amusement
Meditation
Category

ROC Curve:

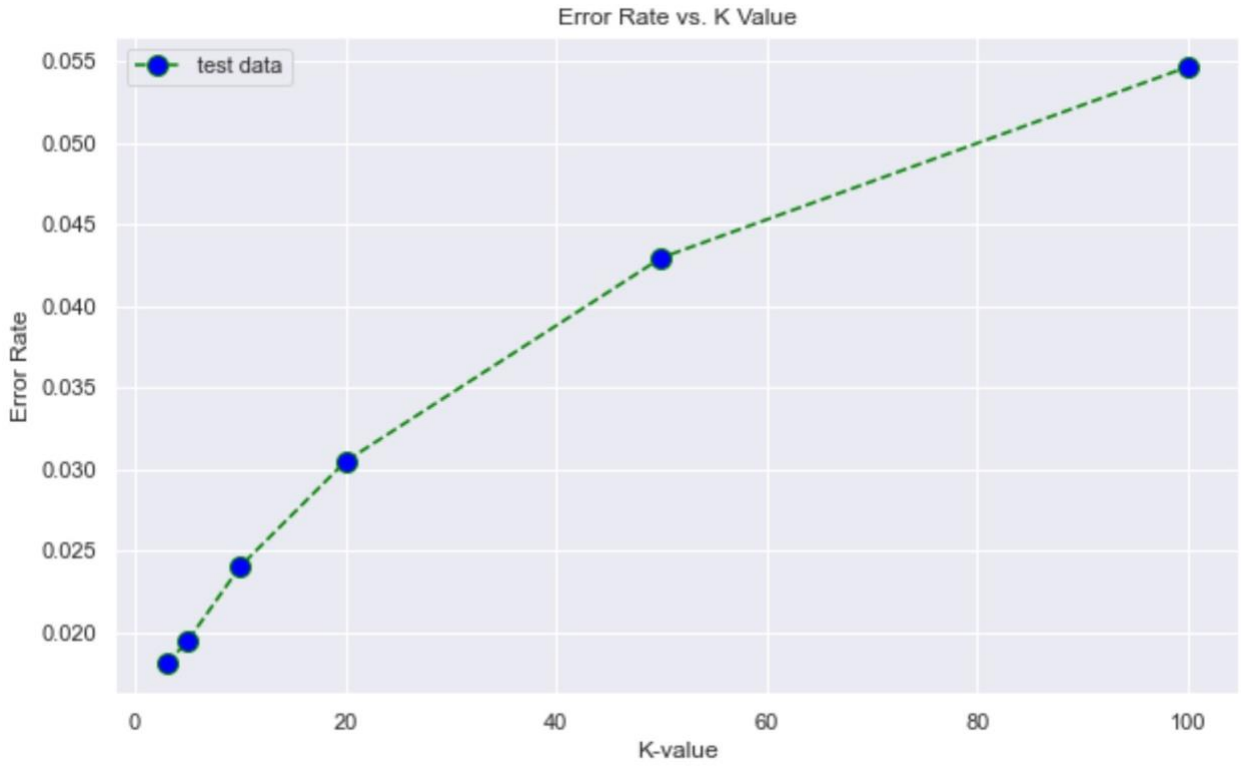


Interpretations:

From the above we observe that:

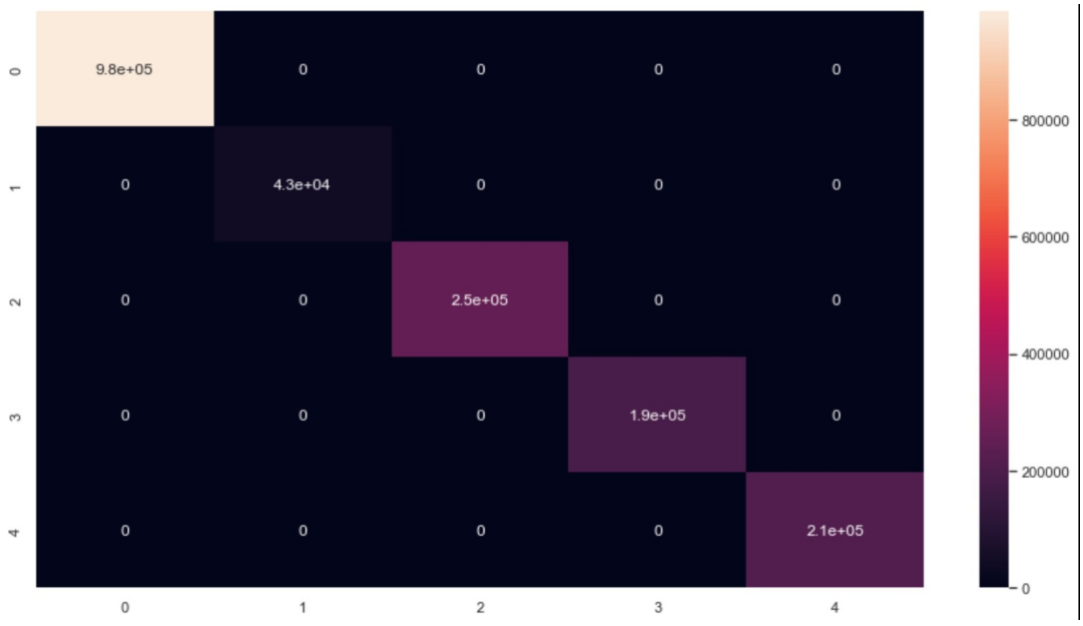
- The KNN Model with K=100 , slightly drops accuracy but still considered as good results, with an accuracy of 94.54%
- The 'Stress' category has the lowest F1 Score, while the Baseline and Meditation have the highest and subsequent F1 scores respectively
- The ROC Curve continues to generate similar results to that of previous k values of all KNN models

FINDING THE RIGHT K VALUE



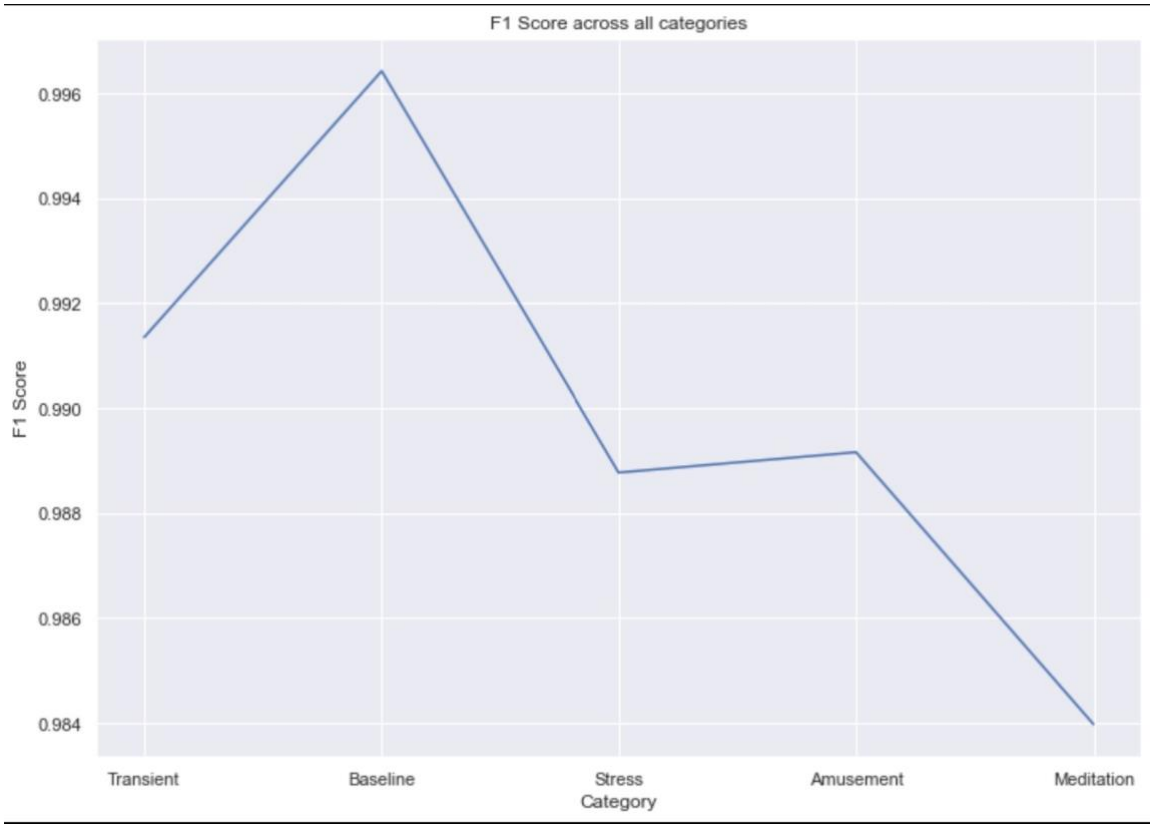
As depicted from the graph, the error rate between k=3 & 5 values is the lowest, and the difference in error rate increases as k value is increased.

RANDOM FOREST

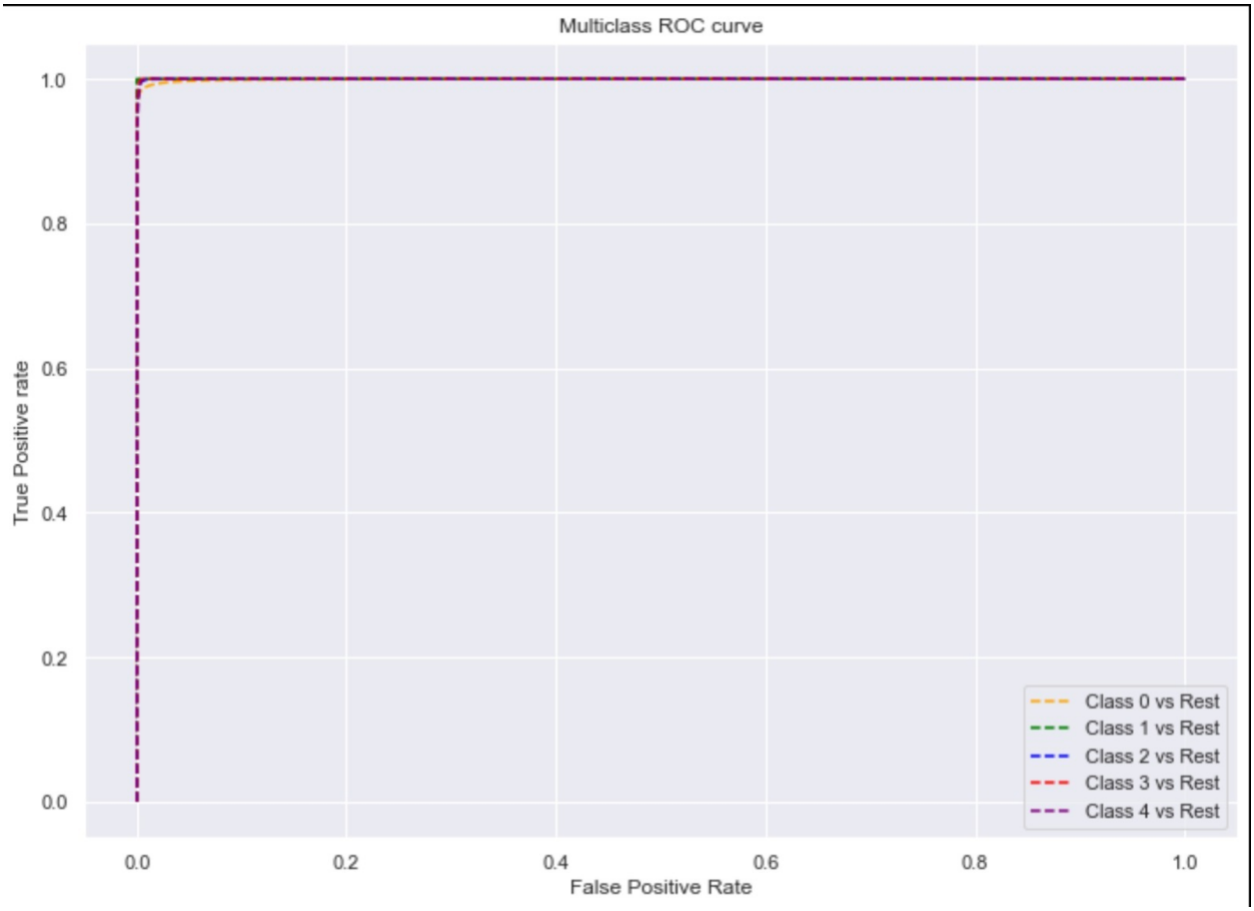


Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9943	0.9885	0.9914	245572
1.0	1.0000	0.9929	0.9964	10807
2.0	0.9858	0.9918	0.9888	63502
3.0	0.9846	0.9938	0.9892	48082
4.0	0.9777	0.9903	0.9840	52129
accuracy			0.9899	420092
macro avg	0.9885	0.9914	0.9899	420092
weighted avg	0.9900	0.9899	0.9899	420092



ROC Curve:

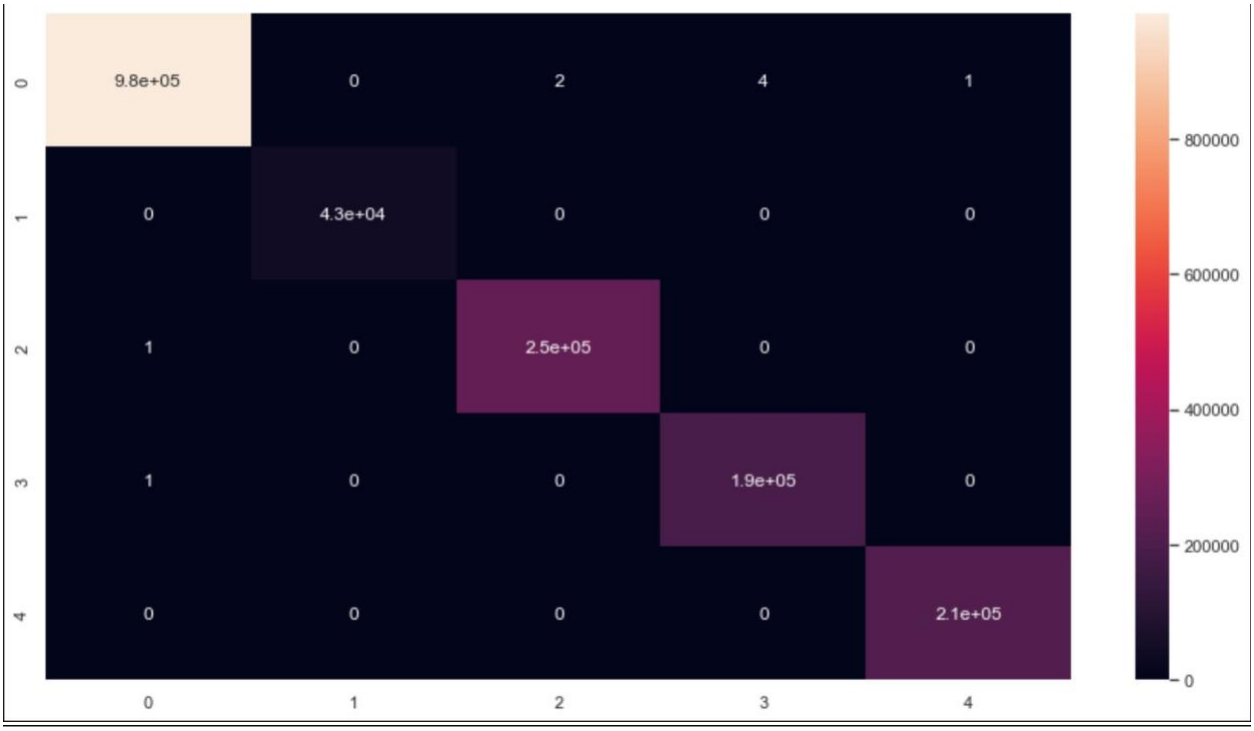


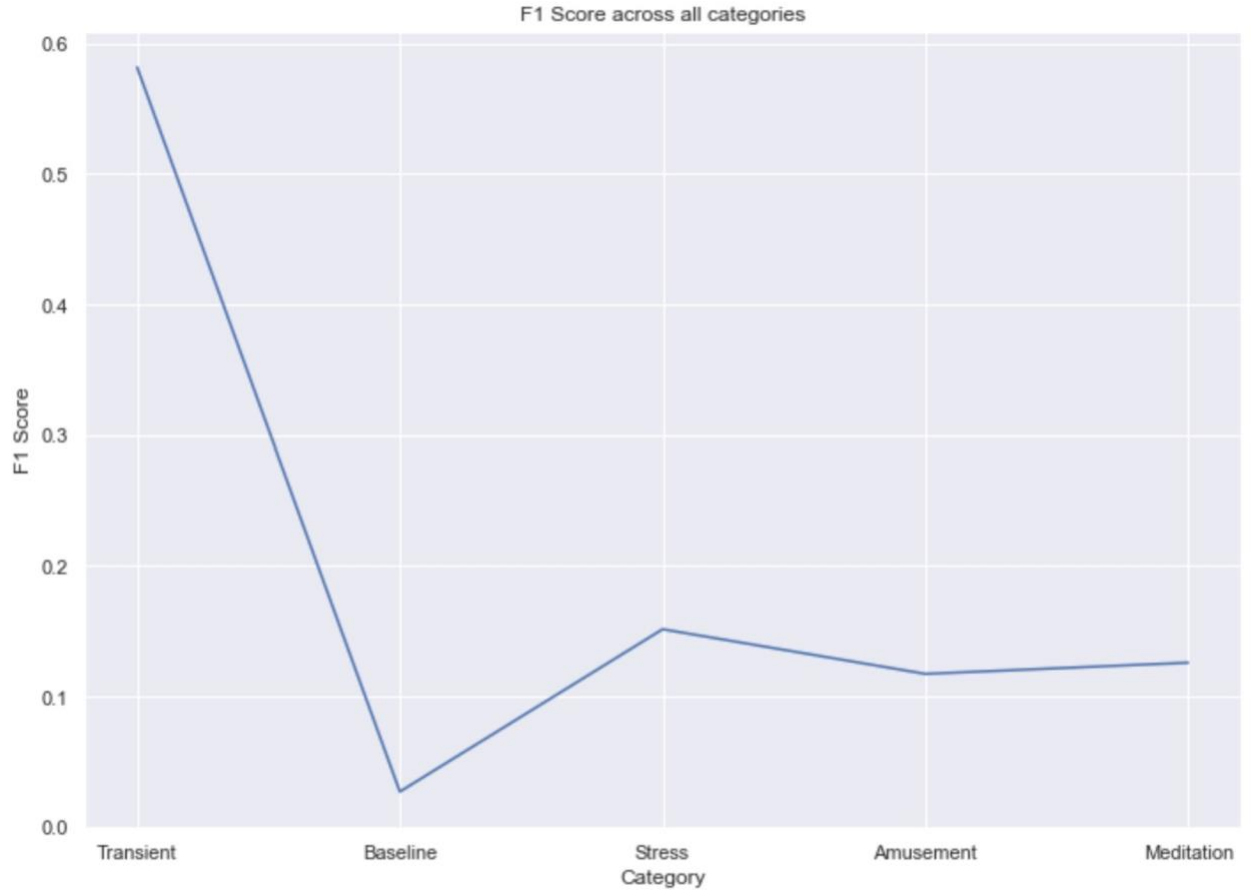
Interpretations:

From the above we observe that:

- The ROC Model shows the best results so far, with an accuracy of 98.99%
- The 'Meditation' category has the lowest F1 Score, while the Baseline has the highest score
- The ROC Curve also gives the perfect results on models so far implemented

RANDOM FOREST ON PCA

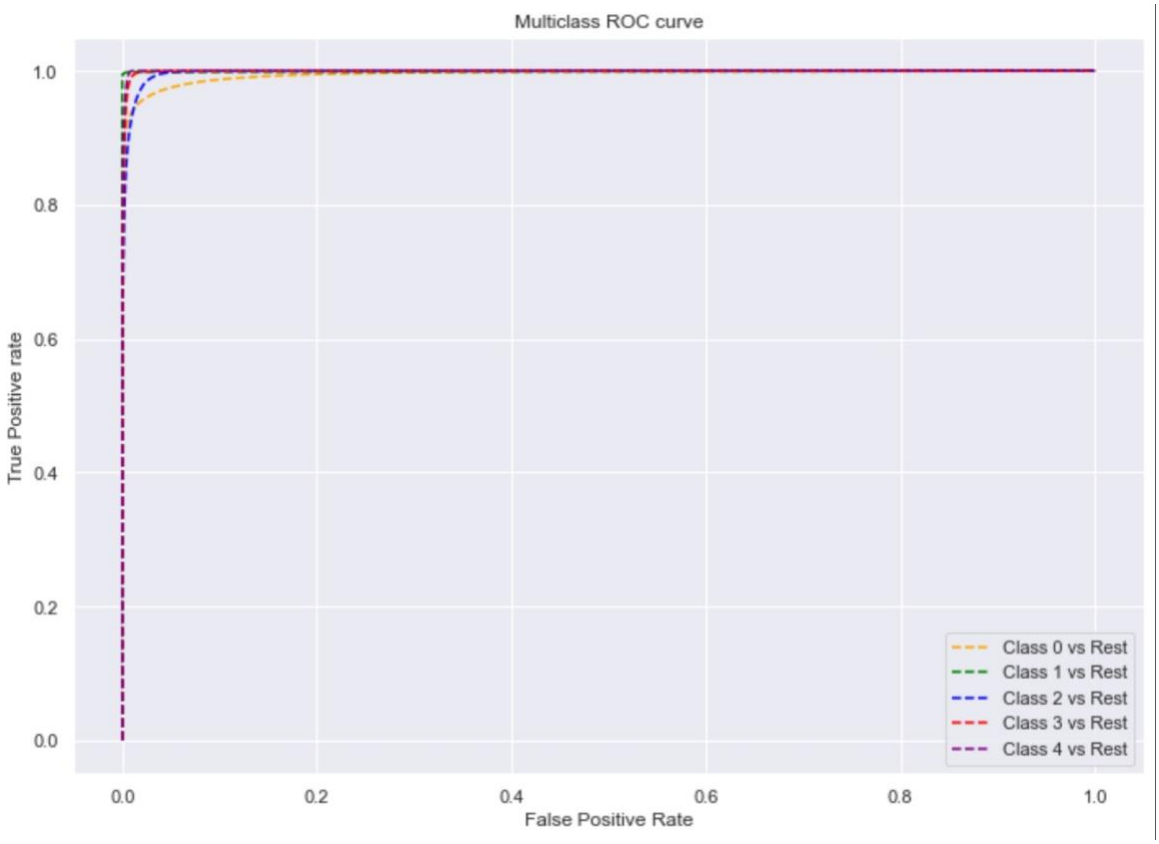




Classification Summary on Test data

	precision	recall	f1-score	support
0.0	0.9765	0.9662	0.9713	245683
1.0	0.9966	0.9865	0.9915	10694
2.0	0.9350	0.9434	0.9392	63532
3.0	0.9533	0.9761	0.9646	47951
4.0	0.9661	0.9843	0.9751	52232
accuracy			0.9667	420092
macro avg	0.9655	0.9713	0.9684	420092
weighted avg	0.9668	0.9667	0.9667	420092

ROC Curve :

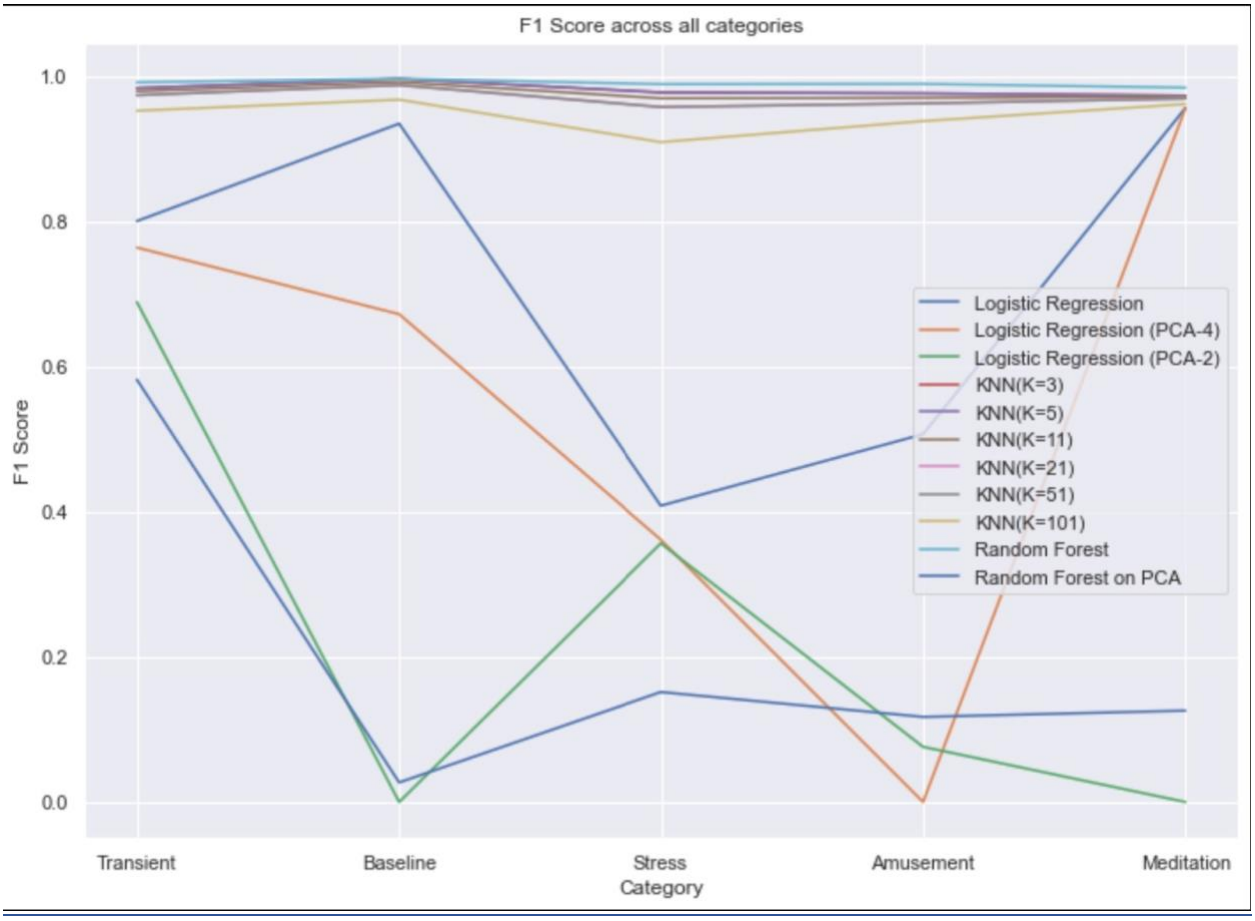


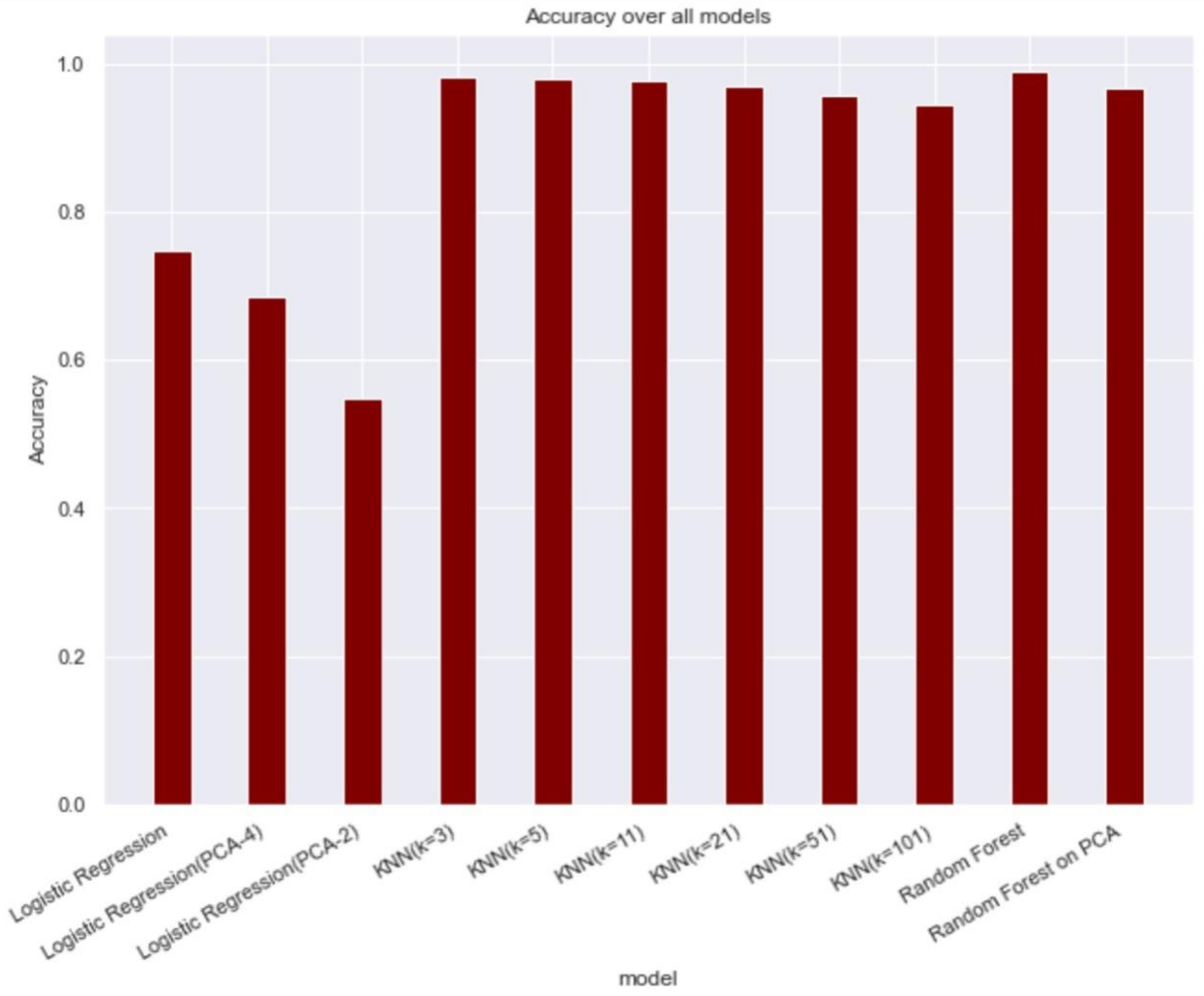
Interpretations:

From the above we observe that:

- The Random Forest Model on PCA also gives great results with an accuracy of 96.67%
- The ‘Transient’ Category has the highest F1 score, while the ‘Baseline’ category has the lowest
- The ROC Curve also looks relatively ideal across the categories.

F1 SCORE ACROSS ALL CATEGORIES:





MODEL IMPLEMENTATION INTERPRETATION:

After implementing the Logistic Regression, KNN Model on different k values and Random Forest Models on the dataset, different parameters for model's selection are calculated. It is observed that while the accuracy and other parameter scores of Logistic Regression aren't relatively great, KNN yields better results on the dataset and Random Forest Model generates the best results on the dataset with an accuracy of 98.99%. The KNN Model is the subsequent best model with k=3 with an accuracy over 98%

Implementation of PCA on Logistic Regression generated bad results with accuracy rates of 68% and 58%. While the PCA implementation worked great on Random Forest giving an accuracy of 96% This means PCA worked best on Random Forest but couldn't improve Logistic Regression model results relatively. Based on the results, the implementation of Random Forest model

generates best results on the dataset once trained, relatively as compared to other models. Logistic Regression is a bad model to implement.

REFERENCES:

1. Dataset: https://archive.ics.uci.edu/ml/datasets/WESAD+%28Wearable+Stress+and+Affect+Detection%29
2. Relevant Papers: https://dl.acm.org/doi/10.1145/3242969.3242985