MINI PROJECT REPORT

Sachin Kumar – B19EE075
Shashi Prakash – B19EE076
Group name – Google Mini

DATASET

We are provided with a large number of images in the given data. We sorted 1500 images for classification having approximately equal number of images from both classes. Then we used matplotlib and NumPy and other online sources to convert these 1500 images into a readable data.

We resized the images to 50*50 and then converted it into a NumPy 1D array . The size of this array for every image is 7500. So, we have a data with 1500 images having 7500 features .

STEPS USED FOR ALL MODELS

1) Data pre-processing : First we pre-processed the given data set and check the null values .Then we converted the categorical data into numerical data. The classes are strings having values ‘with mask’ and ‘without mask’ , which we converted into 1 and 0 respectively.

2)Explanatory Analysis: Mean , standard deviation , minimum , maximum and count of values for every column in dataset is described .

3) Train test split: Before training the model, we split the data set into train and test dataset to check the accuracy of model. The split ratio is 50:50.

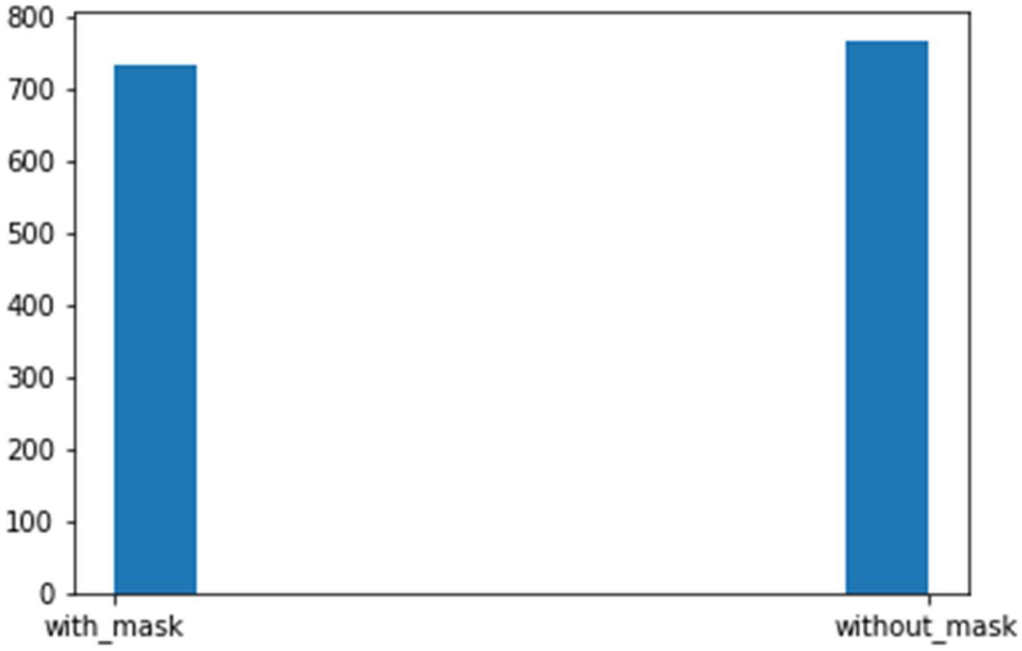
4) Model selection: Then we choose best model for our dataset.

5) Train our model: After that we fit our training data set in the model.

6) Check the accuracy: Then we checked the accuracy , cross validation score and computed confusion matrix and plotted ROC -AUC .

7) Apply PCA : Then we transformed our data into a smaller dimension data using Principle Component Analysis and do the same steps from 3 to 6.

A COUNT - PLOT OF CLASSES



SAMPLE OF IMAGES USED FOR TRAINING MODELS

BRIEF DESCRIPTION OF ALL MODELS

MODEL 1 – LOGISTIC REGRESSION

This model is trained by the above steps mentioned by Sachin (B19EE075) and accuracy of this model is 84.67% and after dimension reduction by PCA its accuracy increased to 85.87%.

MODEL 2 – SVM MODEL

This model is trained by the above steps mentioned by Shashi Prakash (B19EE076) and accuracy of this model is 86% and after dimension reduction by PCA its accuracy decreases to 63.2%.

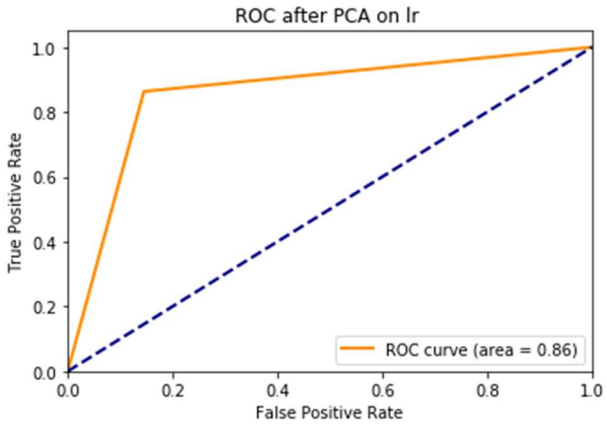
MODEL 3 – MLP MODEL

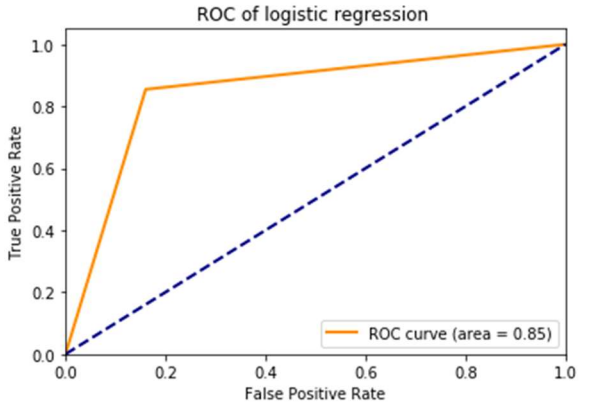
This model is trained by the above steps mentioned by Shashi Prakash (B19EE076) and Sachin Kumar(B19EE075) and accuracy of this model is 87% and after dimension reduction by PCA its accuracy decreases to 83% , so PCA is not beneficial for this data and this model.

COMPARISION TABLE FOR ALL MODELS

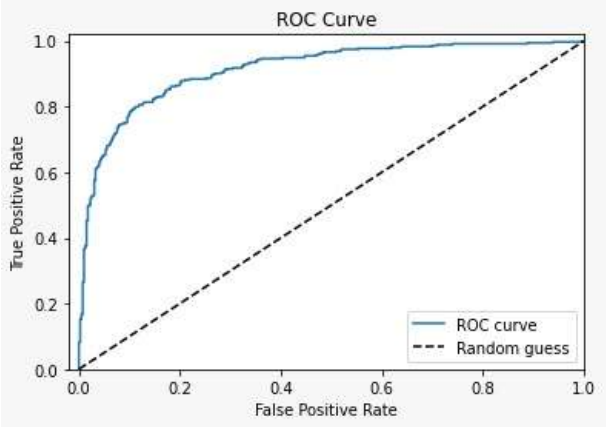
MODEL	ACCURACY	CV – SCORE	CONFUSION MATRIX
LOGISTIC REGRESSION	84.67%	[0.792 0.84 0.84]	[[335 64][51 300]]
LOGISTIC REGRESSION AFTER PCA	85.87%	[0.78 0.88 0.844]	[[341 58] [48 303]]
SVM	86%	[0.87 0.89 0.85]	[[326 70] [34 320]]
SVM AFTER PCA	63.2%	[0.62 0.64 0.68]	[[123 273] [3 351]]
MLP	87%	[0.87 0.89 0.87]	[[335 61] [36 318]]
MLP AFTER PCA	83%	[0.79 0.81 0.81]	[[324 72] [50 304]]

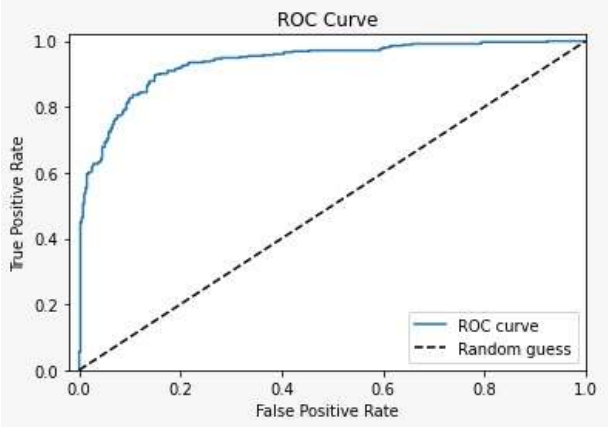
ROC OF LOGISTIC REGRESSION



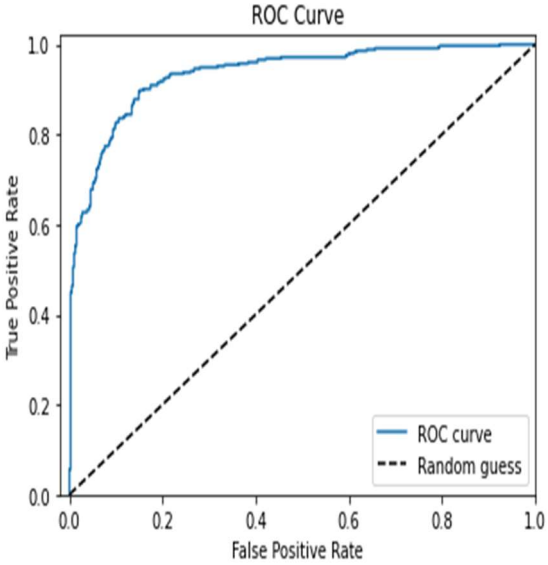


ROC OF SVM

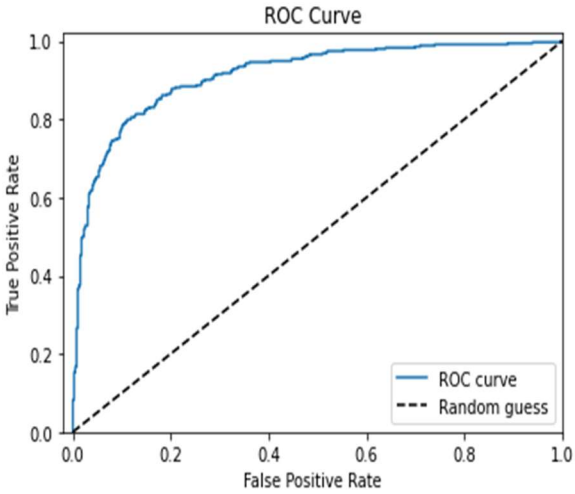




ROC OF MLP



Without PCA



With PCA

CONCLUSION

1. MLP is the best model for this Mask and NO – Mask classification with a good accuracy of 87%.
2. Random Forest classifier with PCA is the fastest method to predict whether a person is wearing mask or not in a given picture .
3. SVM is time consuming and having least accuracy among all models.
4. We also tried this model on our faces , it performed well.