2 PCA

For this problem the data-set that we are going to work with imported from sklearn and its data type is a bunch. We are going to a constraint that only those persons photos will be fetched whose minimum number is 70. We are also going to resize the photos meaning decreasing the amount of pixels

About the data set :

- We have a total of 1288 images with dimension 50 X 37. But for easy operations we will be using 1D form (1080 columns/features) of these matrices.
- Each image has a target (their name) which can be accessed via their categorically encoded number.
- Each number can be used to get their name.
  Ex : 0 refers to George W Bush

2.1 Data Preprocessing

Based on above discussion the parameters of the function fetch_lfw people will be

1. min_faces_per_person = 70, resize = 0.4
2. X (features) being each pixel so we have 1080 columns
3. y (target) the encoded number, in this case the range in 0-6 meaning we have 7 different classes.
4. Now that we have X and y perform test_train_split (train_size = 0.8)

2.2 Eigenfaces Implementation

1. We have to find the mean face by calculating mean along y-axis. Then reshape it accordingly

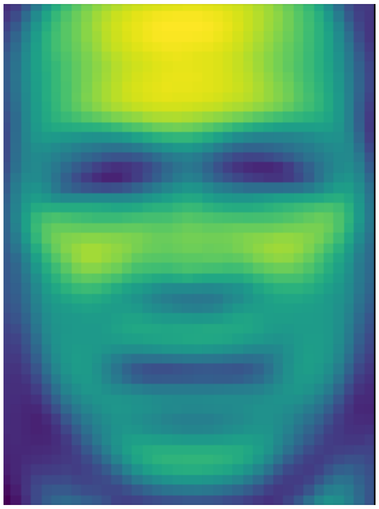
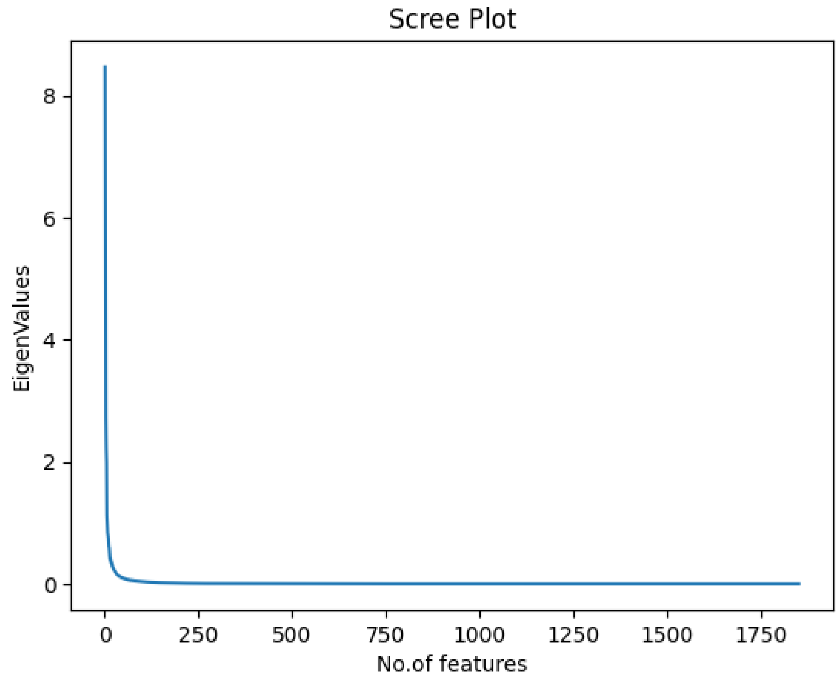
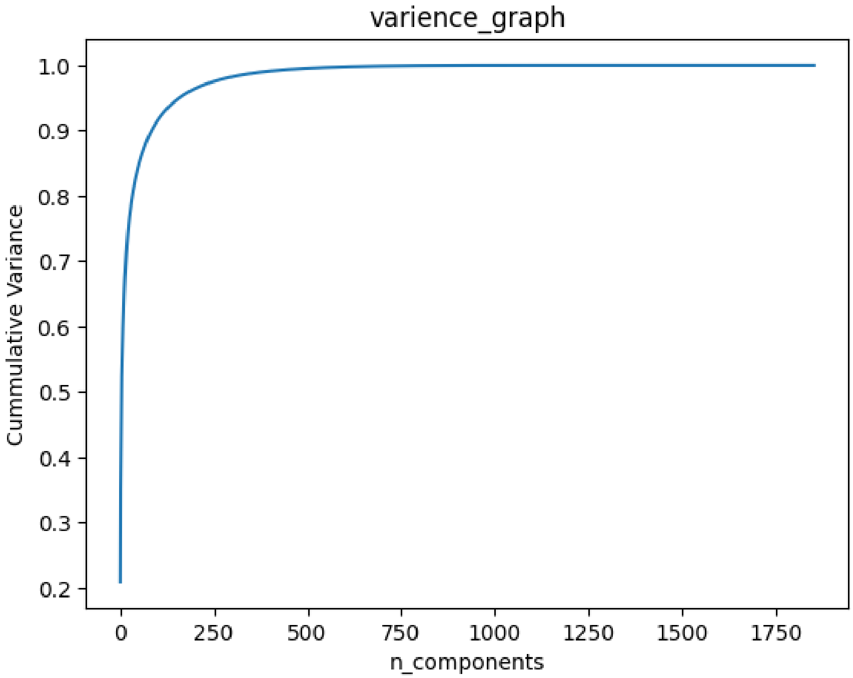


Figure 1: calculated mean face

2. subtract each face with the mean face
3. Get co-variance matrix using inbuilt np.cov() function
4. Get eigen vectors and their corresponding eigen values using np.linalg.eigh()
5. Sort eigen vectors based on eigen values in descending order
6. In eigenvectors 2D list, eigenfaces are arranged column wise meaning each column represents a eigen face
7. Choosing n = 162. For this purpose I have used 3 methods
    - Based on scree plot if we consider only the eigen values with value >= 1 the n will be 3. NOT SUFFICIENT
    - Based on variance plot we I take n = 162 the cumulative variance is 95 which is SUFFICIENT
    - Also based on reconstruction we don't see much of a variation after n=150


Scree Plot
EigenValues
0
2
4
6
8
No.of features
0
250
500
750
1000
1250
1500
1750


varience_graph
Cummulative Variance
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
n_components
0
250
500
750
1000
1250
1500
1750

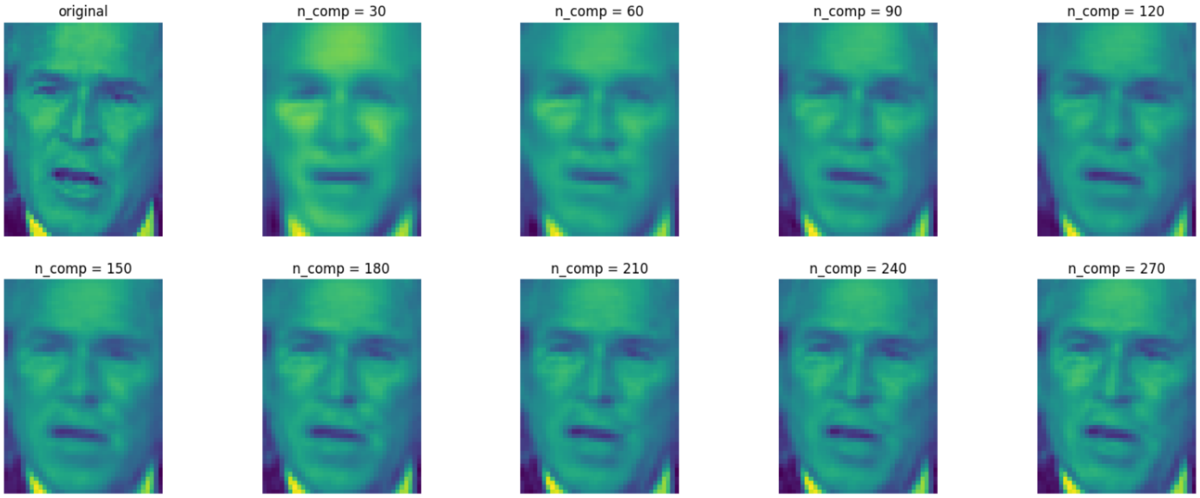


Figure 2: reconstruction

8. Now for feature reduction just multiply the data with projection matrix(matrix with required number of eigen faces as columns)

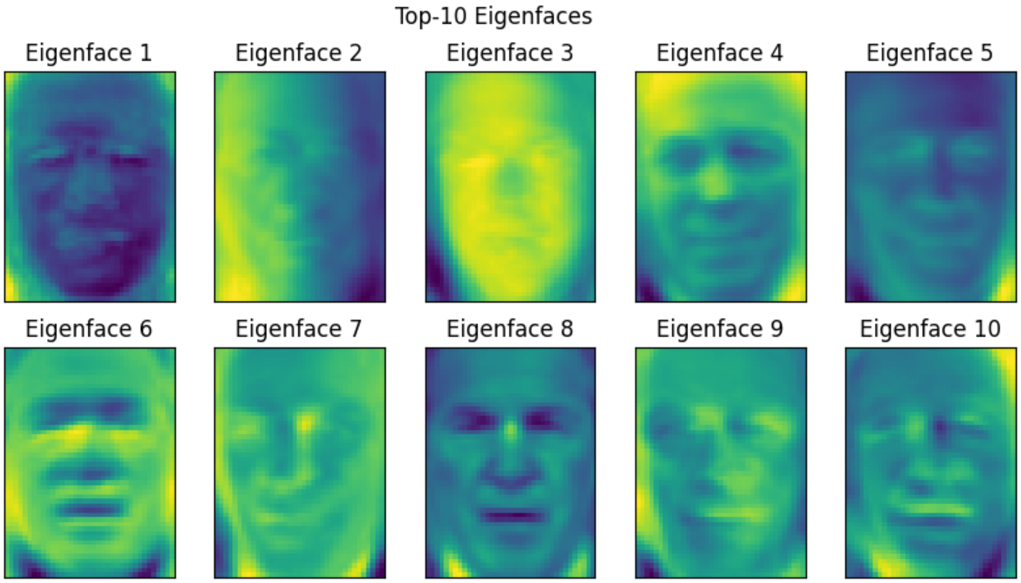


Figure 3: projecting eigen vectors with maximum variance as faces

2.3 Model Training

For Model I have choose three different classifiers KNN, Decision Tree and Random Forest.

2.4 Model Evaluation

Accuracies are as follows :

1. KNN

   Before PCA : 0.5116279069767442
   After PCA : 0.5116279069767442
   Difference : -0.007751937984496138

name	precision	recall	f1-score	support
Ariel Sharon	0.13	0.17	0.15	12
Colin Powell	0.53	0.67	0.59	46
Donald Rumsfeld	0.41	0.39	0.40	23
George W Bush	0.59	0.81	0.68	103
Gerhard Schroeder	0.12	0.04	0.06	26
Hugo Chavez	0.50	0.06	0.11	17
Tony Blair	0.64	0.23	0.33	31

2. Decision Tree

   Before PCA : 0.4883720930232558
   After PCA : 0.4883720930232558
   Difference : 0.08914728682170542

name	precision	recall	f1-score	support
Ariel Sharon	0.08	0.08	0.08	12
Colin Powell	0.31	0.30	0.31	46
Donald Rumsfeld	0.26	0.20	0.23	23
George W Bush	0.61	0.61	0.62	103
Gerhard Schroeder	0.32	0.23	0.27	26
Hugo Chavez	0.17	0.12	0.14	17
Tony Blair	0.28	0.32	0.30	31

3. Random Forest

   Before PCA : 0.5968992248062015
   After PCA : 0.5968992248062015
   Difference : 0.06589147286821695

name	precision	recall	f1-score	support
Ariel Sharon	0.00	0.00	0.00	12
Colin Powell	0.72	0.57	0.63	46
Donald Rumsfeld	1.00	0.17	0.30	23
George W Bush	0.49	0.96	0.65	103
Gerhard Schroeder	0.33	0.04	0.07	26
Hugo Chavez	0.00	0.00	0.00	17
Tony Blair	0.50	0.23	0.31	31

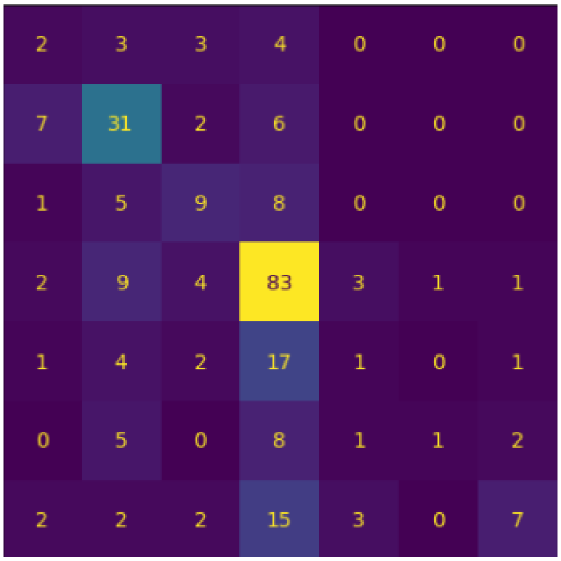
   Although it is expected that accuracy should decrease after PCA as data is lost it is completely possible that the accuracy of the model is better (Difference is -ve) after PCA reasons being :

   - removes noise
   - removes irrelevant information

Reasons For Misclassification

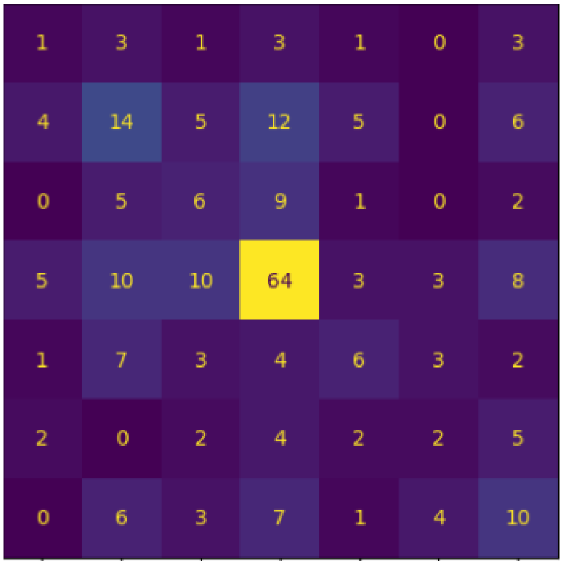
The count of George W-Bush is 530 which is like 52% of data More Weight and for Hugo Chavez it is only 7% Less weights

- Overfitting on George W-Bush
- Underfitting on Hugo Chavez



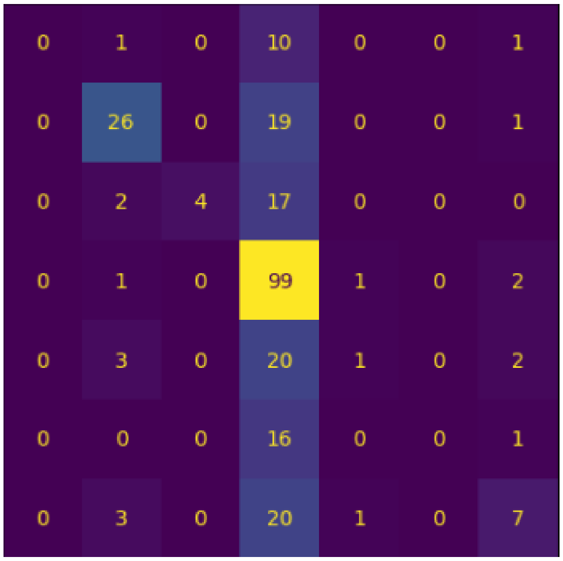
2	3	3	4	0	0	0
7	31	2	6	0	0	0
1	5	9	8	0	0	0
2	9	4	83	3	1	1
1	4	2	17	1	0	1
0	5	0	8	1	1	2
2	2	2	15	3	0	7

Figure 4: KNN



1	3	1	3	1	0	3
4	14	5	12	5	0	6
0	5	6	9	1	0	2
5	10	10	64	3	3	8
1	7	3	4	6	3	2
2	0	2	4	2	2	5
0	6	3	7	1	4	10

Figure 5: Decision Tree



0	1	0	10	0	0	1
0	26	0	19	0	0	1
0	2	4	17	0	0	0
0	1	0	99	1	0	2
0	3	0	20	1	0	2
0	0	0	16	0	0	1
0	3	0	20	1	0	7

Figure 6: Random Forest

The images provided are not uniform, meaning :

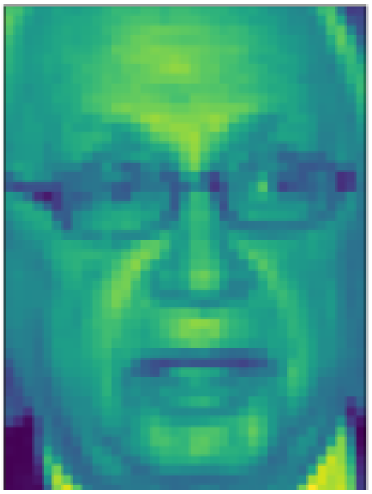


Figure 7: Few images have glasses

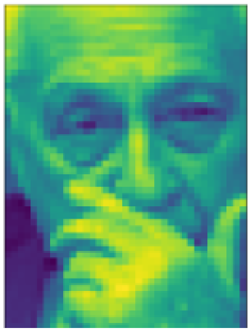


Figure 8: Face is covered by hands

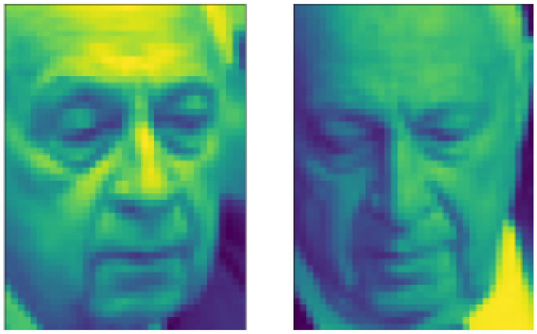


Figure 9: Eyes are closed

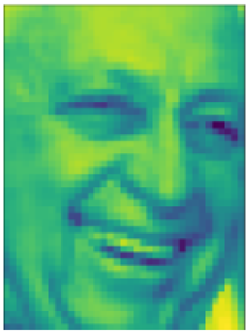


Figure 10: Teeth is visible

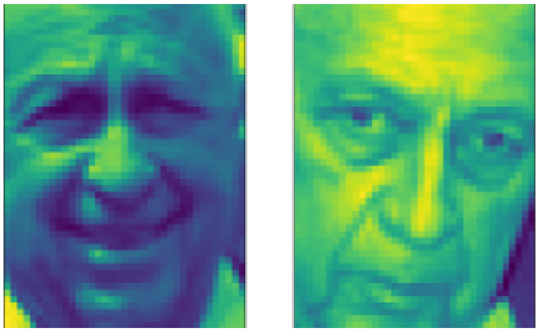


Figure 11: Few are way darker and other more brighter (CONTRAST)

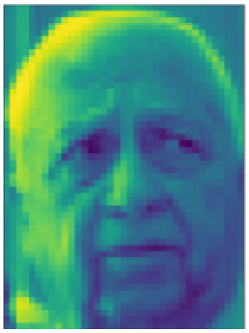


Figure 12: Entire head is visible

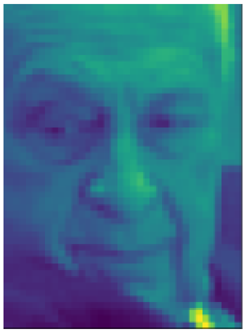


Figure 13: Clothes are visible

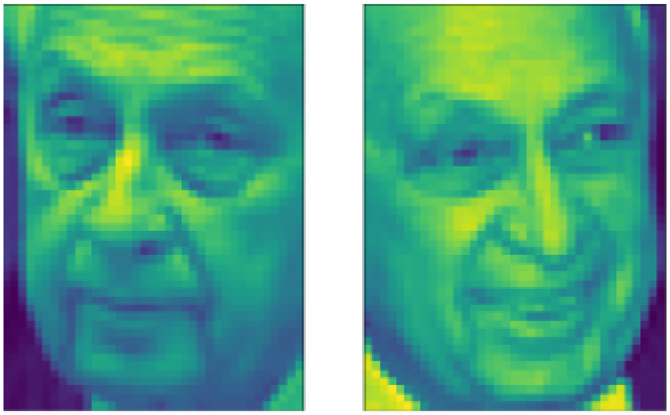


Figure 14: Some are left dominant and other right dominant poses

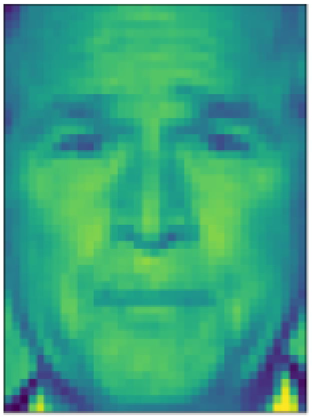
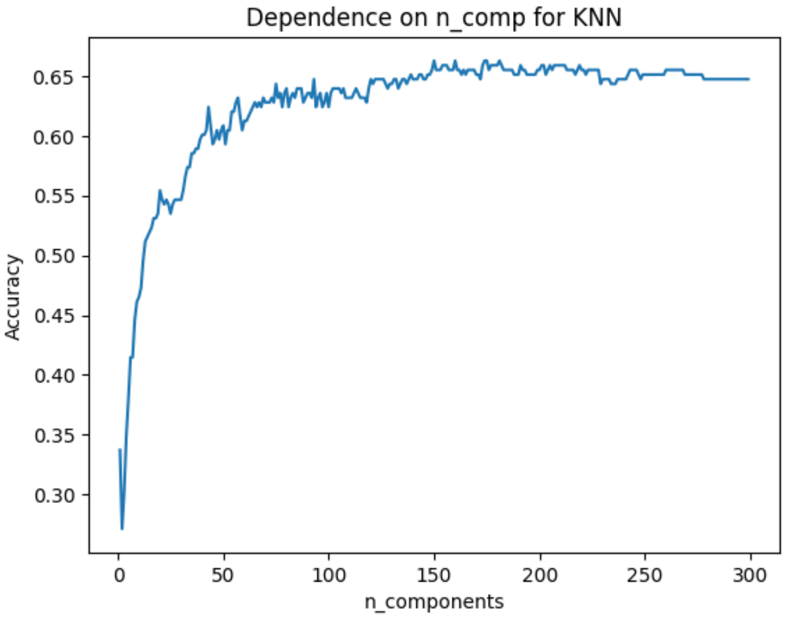
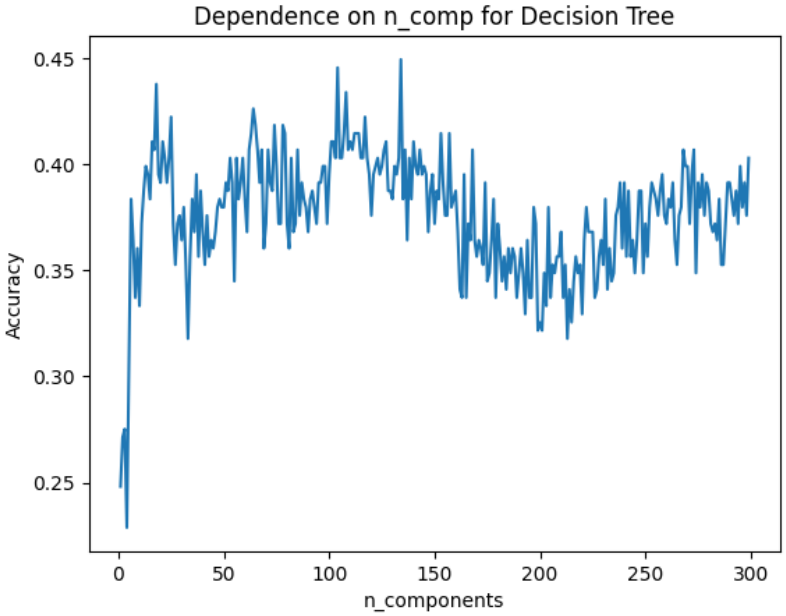
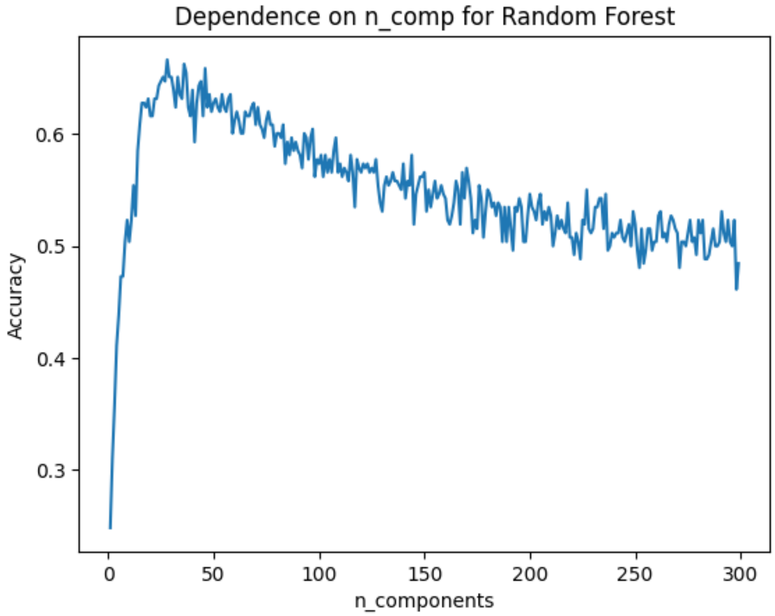


Figure 15: In some lips are closed

2.5 Experiment with different values of n_components






Dependence on n_comp for Random Forest
Accuracy
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
n_components

.