Code ▾

# Machine Learning Exercises

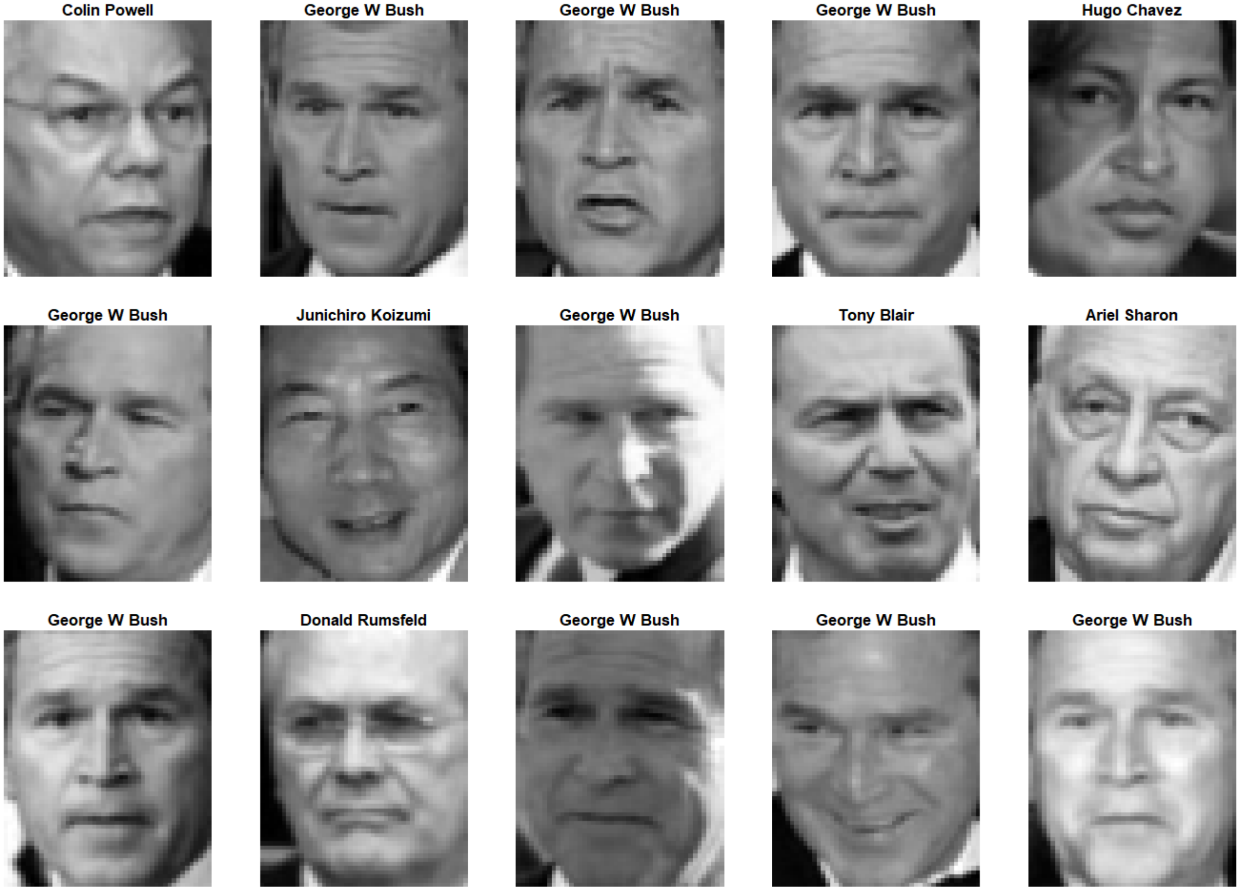
10/07/2021

Code

## 1. The faces in the wild data - face recognition

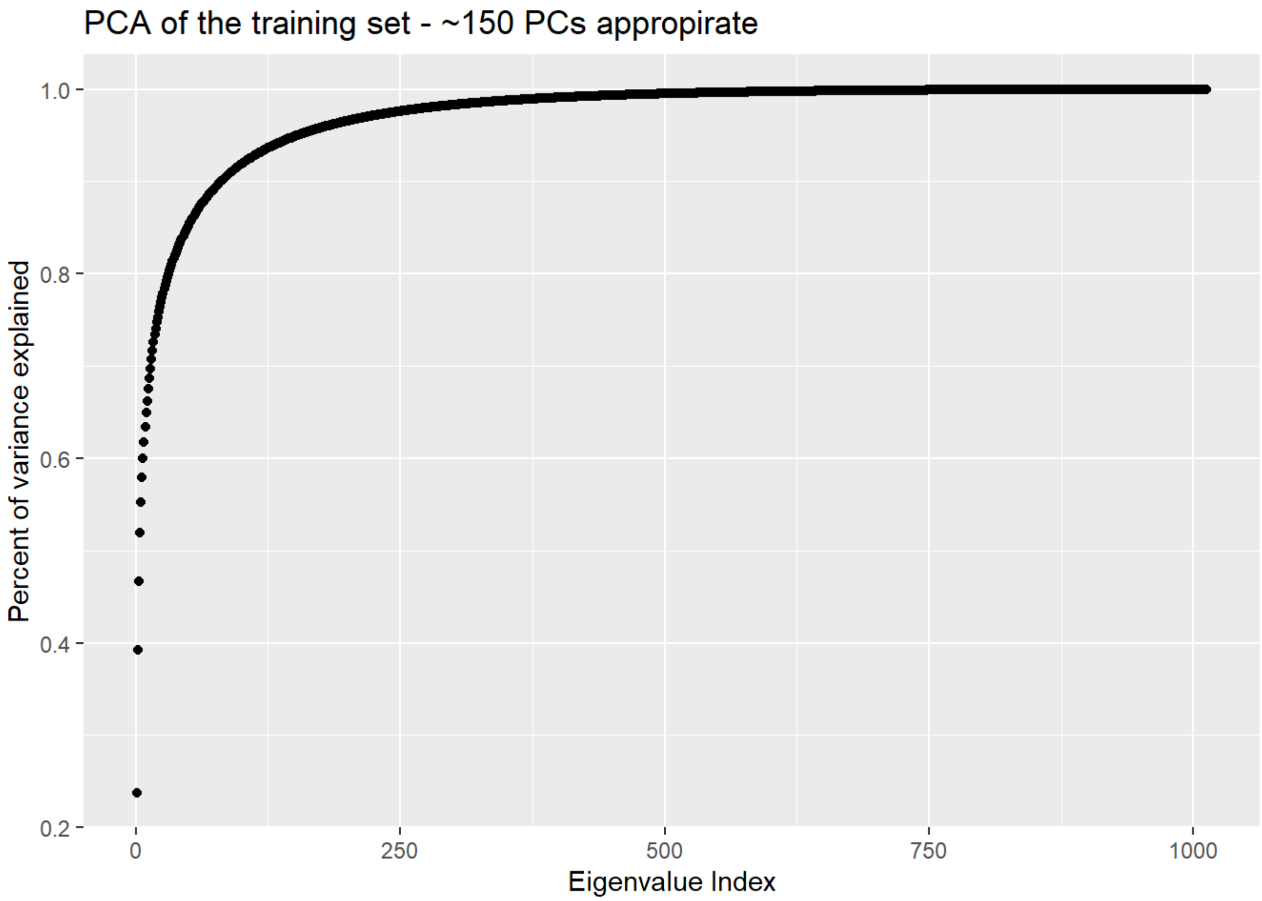
Data from scikit-learn package. Persons with more than 60 face figures were extracted. The following is the first few records in the data.

Code



Colin Powell
George W Bush
George W Bush
George W Bush
Hugo Chavez
George W Bush
Junichiro Koizumi
George W Bush
Tony Blair
Ariel Sharon
George W Bush
Donald Rumsfeld
George W Bush
George W Bush
George W Bush

Code



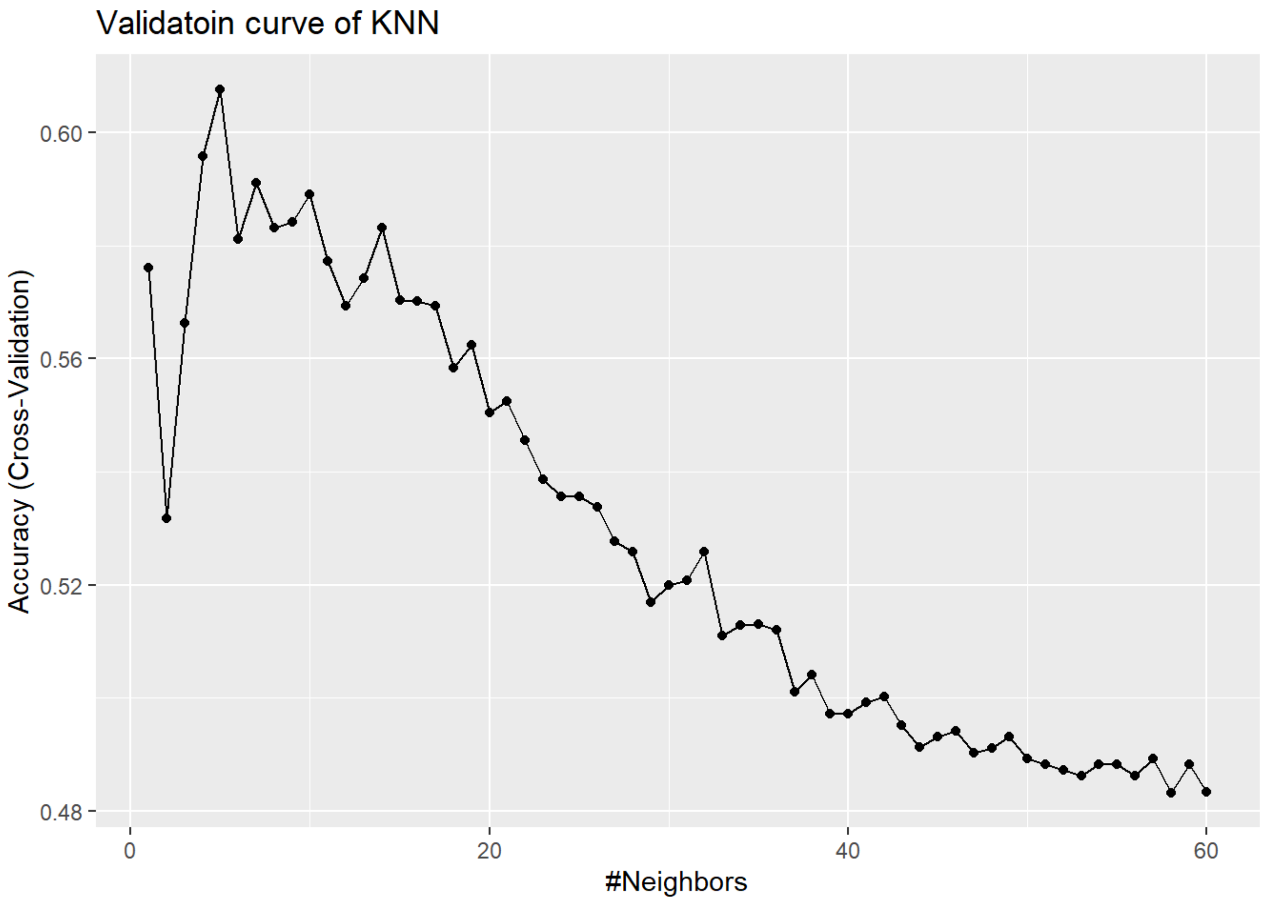
Code

Each image contains 2914 pixels, and so use PCA to reduce dimension. A PCA analysis is as shown above. And, ~150 PCs should suffice.

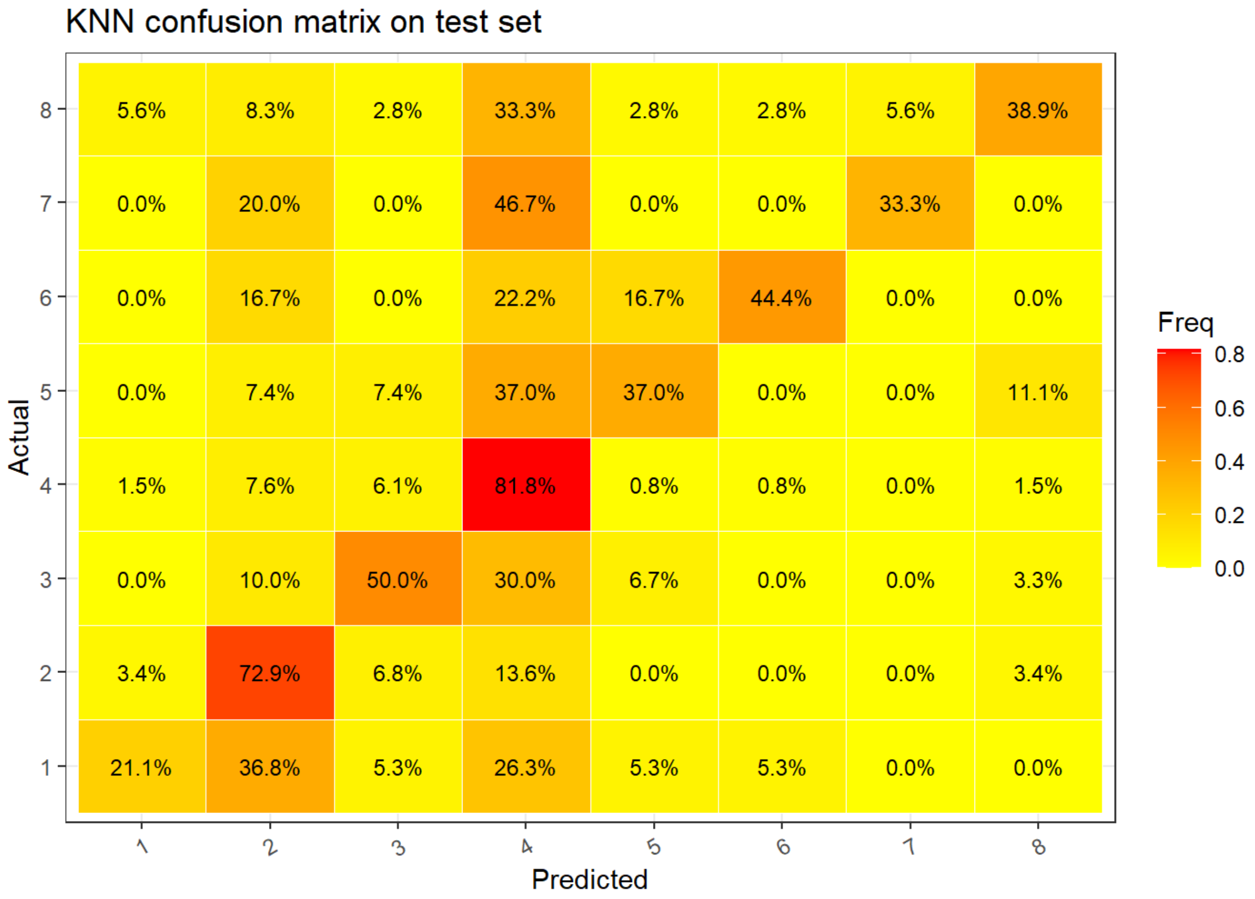
# 2. The KNN model

## 2.1 The regular KNN

Code


Validatoin curve of KNN
Accuracy (Cross-Validation)
0.60
0.56
0.52
0.48
0
20
40
60
#Neighbors

Code


KNN confusion matrix on test set
Actual
8 5.6% 8.3% 2.8% 33.3% 2.8% 2.8% 5.6% 38.9%
7 0.0% 20.0% 0.0% 46.7% 0.0% 0.0% 33.3% 0.0%
6 0.0% 16.7% 0.0% 22.2% 16.7% 44.4% 0.0% 0.0%
5 0.0% 7.4% 7.4% 37.0% 37.0% 0.0% 0.0% 11.1%
4 1.5% 7.6% 6.1% 81.8% 0.8% 0.8% 0.0% 1.5%
3 0.0% 10.0% 50.0% 30.0% 6.7% 0.0% 0.0% 3.3%
2 3.4% 72.9% 6.8% 13.6% 0.0% 0.0% 0.0% 3.4%
1 21.1% 36.8% 5.3% 26.3% 5.3% 5.3% 0.0% 0.0%
1 2 3 4 5 6 7 8
Predicted
Freq
0.8
0.6
0.4
0.2
0.0

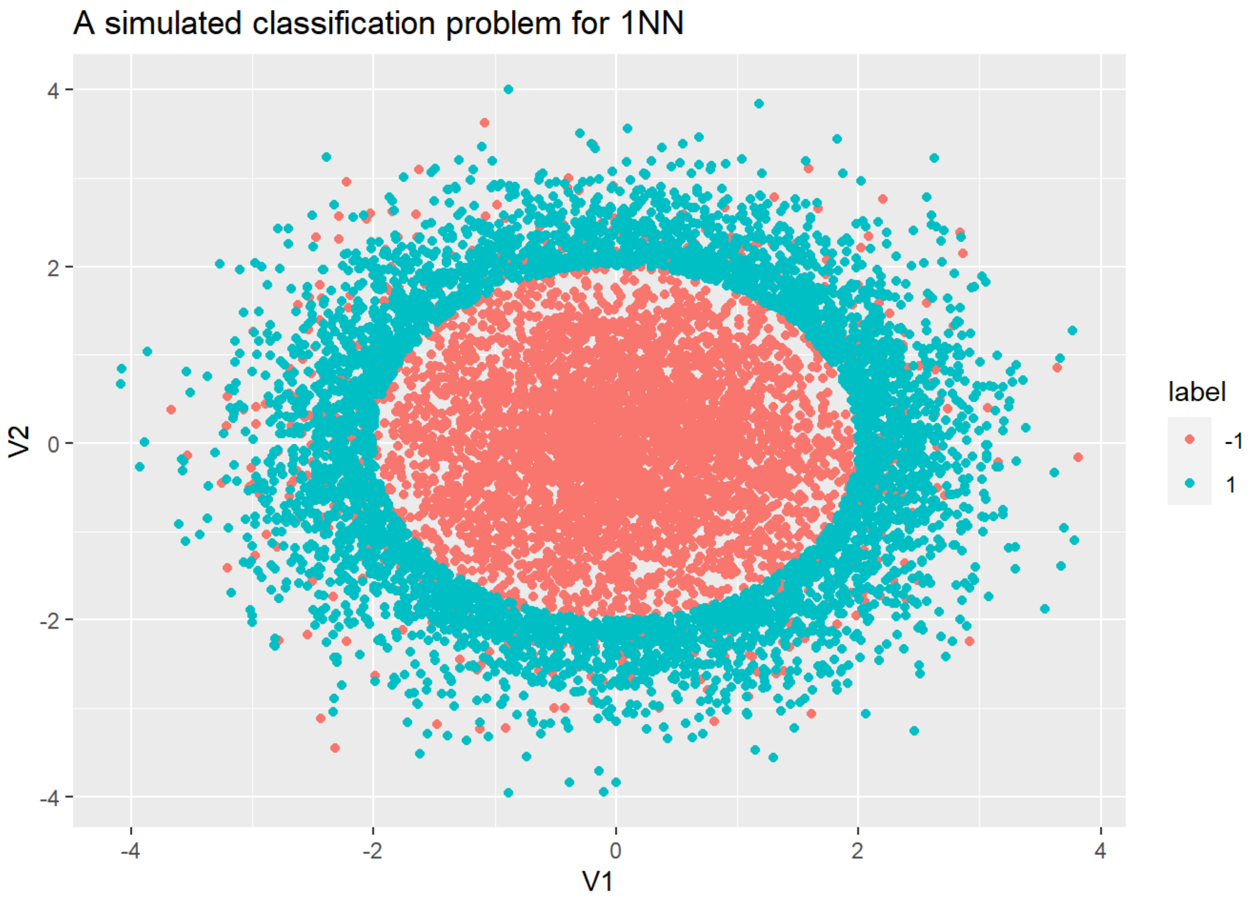
The accuracy of KNN is: 61.61%.

## 2.2 1NN to gain information about Bayes error

### 2.2.1 1NN to gain information about Bayes error - Simulated data

*Asymptotically, 1NN error is never more than twice the Bayes error.*

Code



Code

Code

The 1NN error on the test data is: 14.32%. So, the Bayes error must be more than 7.16%. The theoretical Bayes error (calculated from this trial) is 7.30%. Nonetheless, in practice, the asymptotic part is always in doubt.

### 2.2.2 1NN to gain information about Bayes error - face data

Code

The 1NN error on the test data is: 36.61%. So, the Bayes error must be more than 18.30%.

## 2.3 The Nested Cross Validation (NCV)

CV error is too optimistic. NCV is used to assess model error and stability. The following are NCV for KNN and SVM models using simulated data as earlier but with different sizes.

Code

Nested CV for the KNN model

| trial | accuracy_test_ncv |
|---|---|
| 1 | 0.91 |
| 2 | 0.90 |
| 3 | 0.91 |
| 4 | 0.92 |
| 5 | 0.91 |
| 6 | 0.91 |
| 7 | 0.92 |
| 8 | 0.93 |
| 9 | 0.91 |
| 10 | 0.93 |

Code

Nested CV for the SVM model

| trial | accuracy_test_ncv |
|---|---|
| 1 | 0.90 |
| 2 | 0.91 |
| 3 | 0.91 |
| 4 | 0.91 |
| 5 | 0.90 |

### 2.4 Naive Bayes

Code

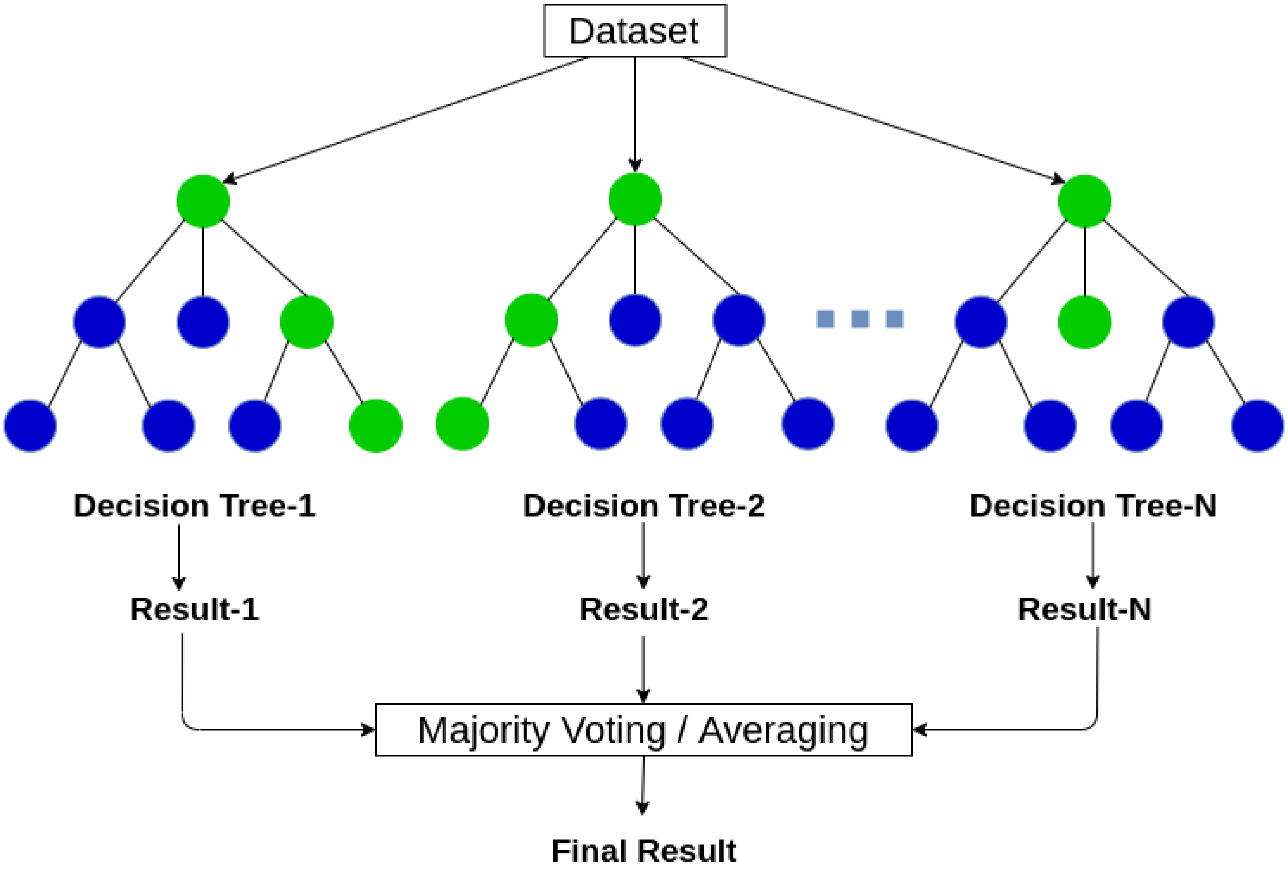
The accuracy of KNN is: 65.18%. It performs better than KNN as well as random forest.

## 3. Random Forest

**The Random Forest:**

- Unlike a single decision tree, which is notorious for its variability, random forest builds a multitude of trees.

- The forest makes a de1cision based on a collective majority vote of all the individual trees in a classification problem. The joint effort helps to reduce variance brought about by a single decision tree.
- Just a portion of the predictors are randomly selected to split a parent node, and this will help dampen the effect of the predominant predictors, decreasing variability in prediction.
- Lower tendency towards over-fitting.
- Small number of hyper-parameters to tune compared with Boosting/XGBoost.
- Random forests use out of bag sample and error (OOB error) to measure out of sample performance, and the favorable aspect is that the OOB errors can be produced during the training process.



Random Forecast: Click to go to chart source (https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/)

## 3.1 Hyper-parameter tuning and model fitting

At the lowest OOB error, the best parameters are:

Code

Best Parameters for the Random Forest Model

| mtry | OOBError | No_trees | Node_size |
|---|---|---|---|
| 63 | 0.34 | 400 | 3 |

Code

## 3.2 The Random Forest Model Performance

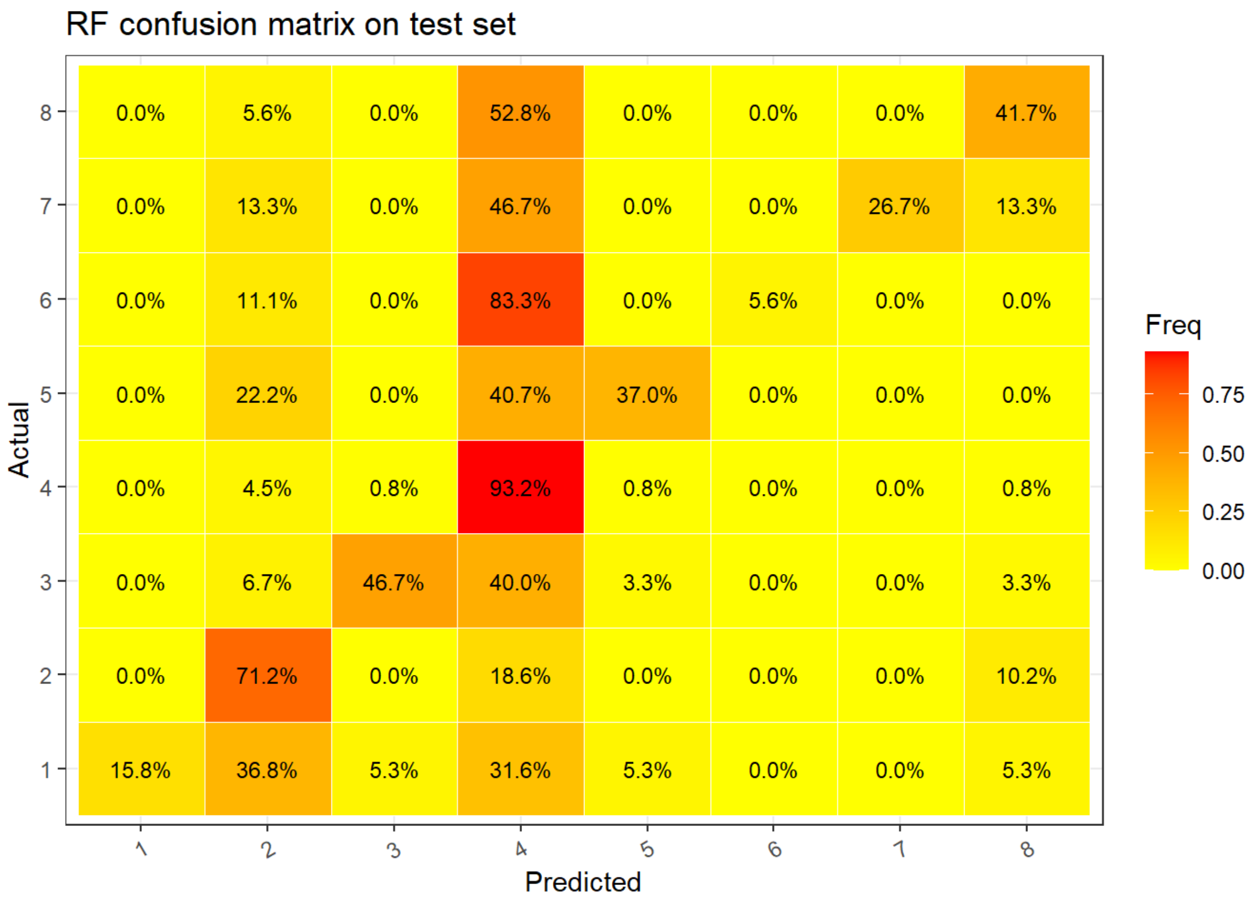
### 3.2.1 The Confusion Matrix

Code

The confusion matrix can show the model prediction vs the actual:

### 3.2.2 The notch difference plot

Code



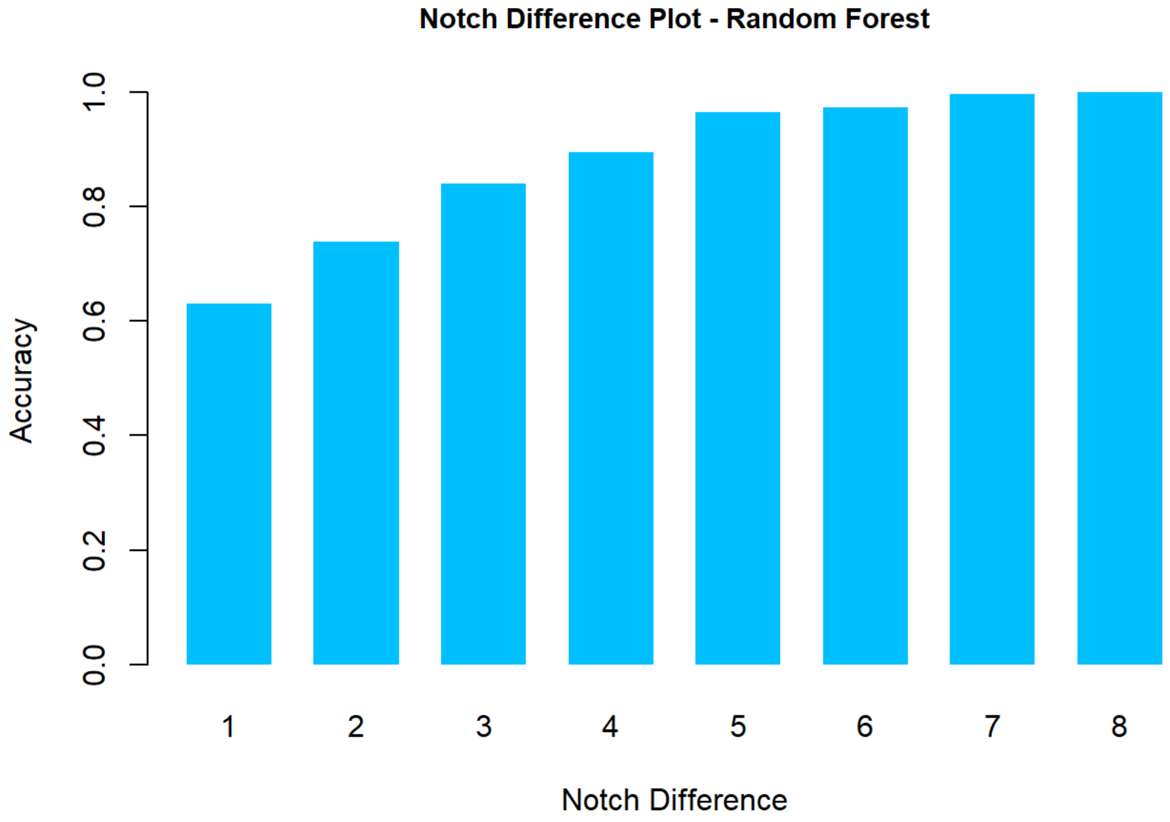
There are totally 336 obligors in the test data set. Model prediction accuracy in test data is: 63.10%

Code

Notch difference table - Random Forest

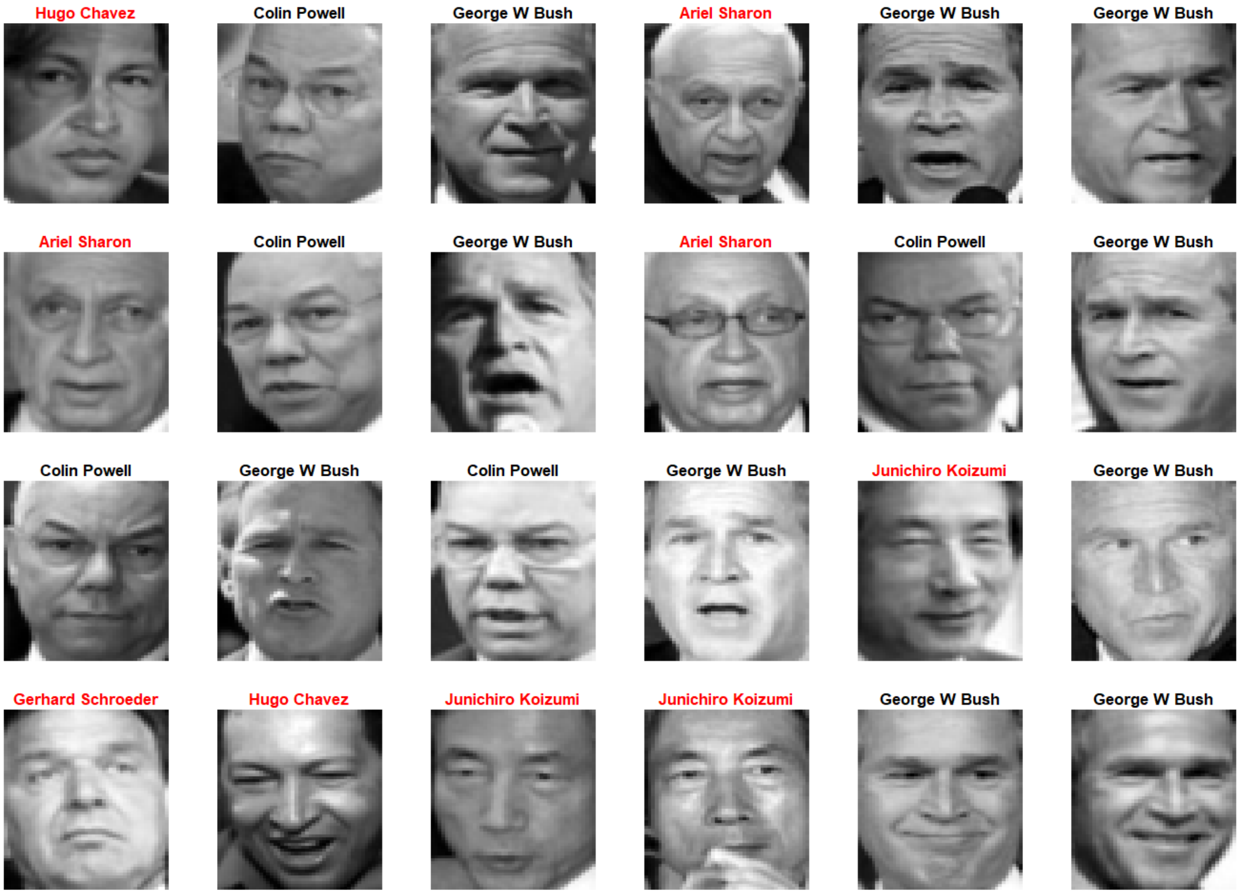
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 0.631 | 0.7381 | 0.8393 | 0.8958 | 0.9643 | 0.9732 | 0.997 | 1 |

Code


Notch Difference Plot - Random Forest

### 3.2.3 The predicted vs actual images

Code



Code

## 4. The Support Vector Machine (SVM)

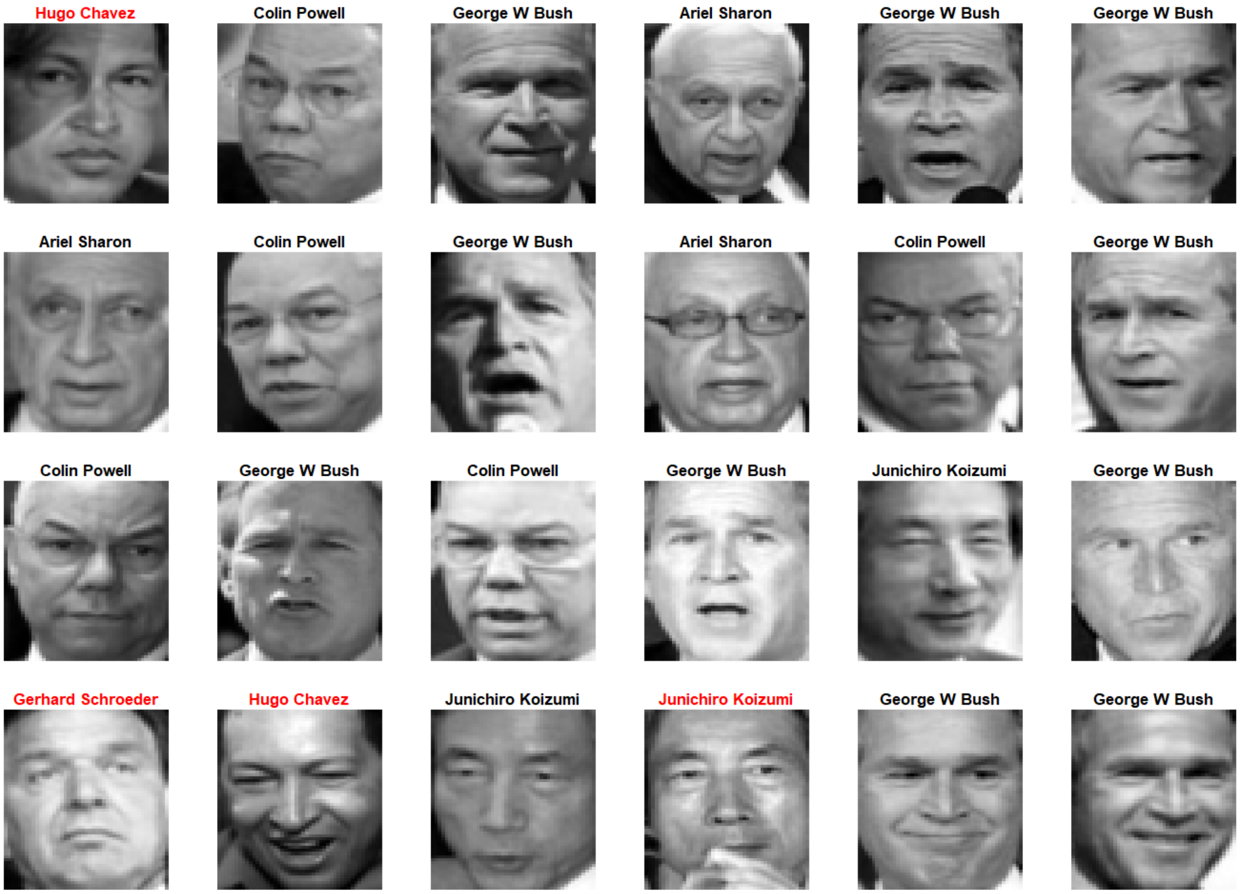
SVM improves substantially from the previous two learners.

Code

Code

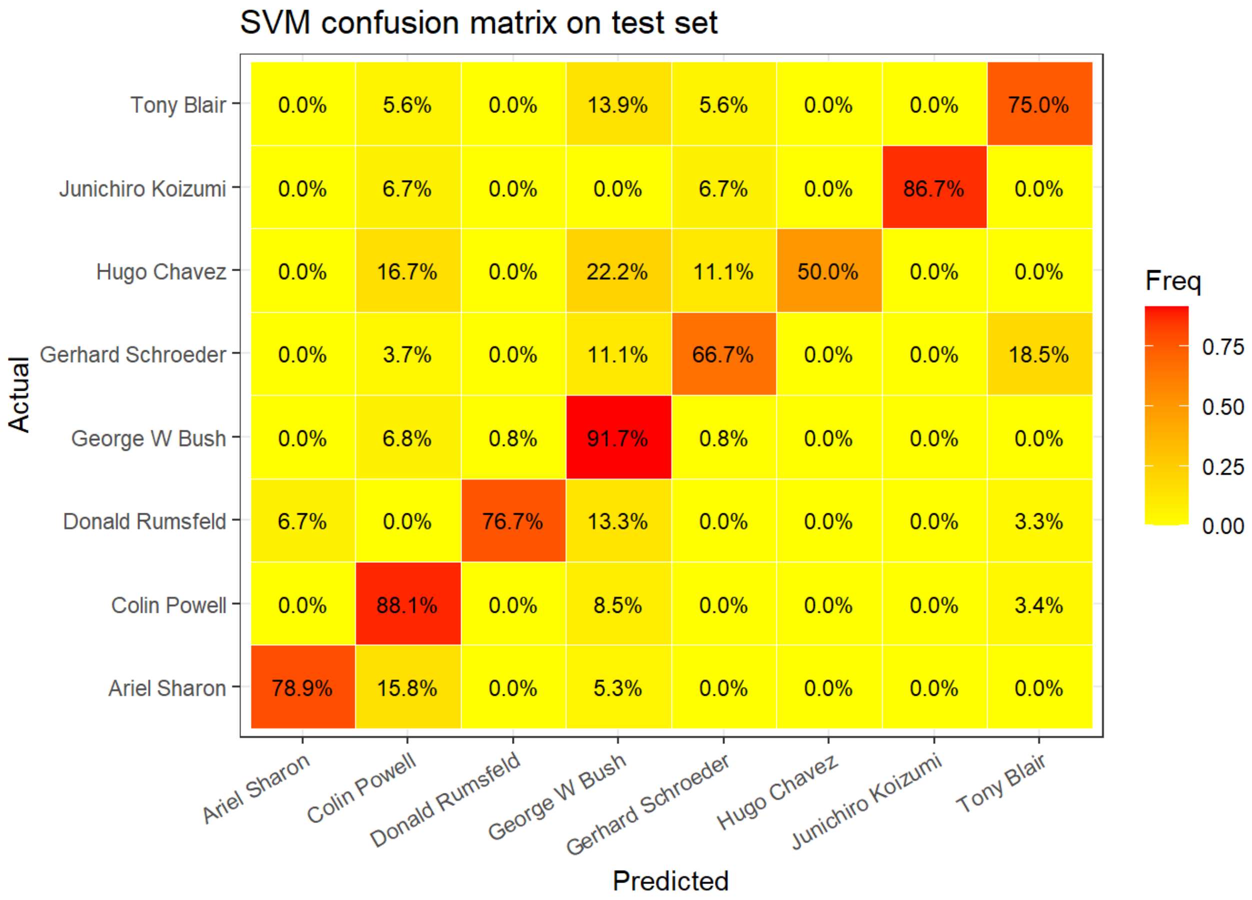
The accuracy of SVM is: 82.74%. The following shows the prediction among the first 24 images. Images labeled in red are miss-classified.

Code


Hugo Chavez
Colin Powell
George W Bush
Ariel Sharon
George W Bush
George W Bush
Ariel Sharon
Colin Powell
George W Bush
Ariel Sharon
Colin Powell
George W Bush
Colin Powell
George W Bush
Colin Powell
George W Bush
Junichiro Koizumi
George W Bush
Gerhard Schroeder
Hugo Chavez
Junichiro Koizumi
Junichiro Koizumi
George W Bush
George W Bush

Code

Code


SVM confusion matrix on test set
Actual
Tony Blair 0.0% 5.6% 0.0% 13.9% 5.6% 0.0% 0.0% 75.0%
Junichiro Koizumi 0.0% 6.7% 0.0% 0.0% 6.7% 0.0% 86.7% 0.0%
Hugo Chavez 0.0% 16.7% 0.0% 22.2% 11.1% 50.0% 0.0% 0.0%
Gerhard Schroeder 0.0% 3.7% 0.0% 11.1% 66.7% 0.0% 0.0% 18.5%
George W Bush 0.0% 6.8% 0.8% 91.7% 0.8% 0.0% 0.0% 0.0%
Donald Rumsfeld 6.7% 0.0% 76.7% 13.3% 0.0% 0.0% 0.0% 3.3%
Colin Powell 0.0% 88.1% 0.0% 8.5% 0.0% 0.0% 0.0% 3.4%
Ariel Sharon 78.9% 15.8% 0.0% 5.3% 0.0% 0.0% 0.0% 0.0%
Ariel Sharon
Colin Powell
Donald Rumsfeld
George W Bush
Gerhard Schroeder
Hugo Chavez
Junichiro Koizumi
Tony Blair
Predicted
Freq
0.75
0.50
0.25
0.00