# MiniProject 1

In this mini-project, the concepts of Principal Component Analysis, Naive Bayes and Classification with Gradient Descent will be put to application. You will be given a small dataset containing some face images of some people. Firstly, you will perform principal component analysis of all images combined. That will be used to get a small number of principal eigen vectors to convert each image to a reduced vector. Then you will learn a classifier using naive bayes method a linear classifier using gradient descent. You will submit your codes and a report pdf in your final submission. You will use python2 for this mini-project.

## Dataset

- You are provided a subset of original dataset with labels. `<a>_<b>.jpg` image is bth image of ath class.
- You will use just images for Part-1 and prepare your report.
- For Part-2 and Part-3, you can use images and labels while developing code. This is just practice dataset so don't assume anything from it. We will test your code for Part-2 and Part-3 on different splits of original dataset. Also images names may not be same like this given dataset.

## Part-1 : Principal Component Analysis

- Perform PCA over all images. Get the co-efficients of all components for all images.
- We can reconstruct the image back using a small number of components. Plot a graph showing the total mean square error over all train images vs the number of principal components used to reconstruct. Include this plot in your submission. (Use a reasonable range for number of components)
- Decide $N$, the number of principal components required such that the reconstructed images will have mean squared error less than 20% over all train images. Display those $N$ principal components as reconstructed images. You will see some base structures of the faces. Include these images in your report.
- Use scatterplots to examine how the images are clustered in the 1D, 2D and 3D space using the required number of principal components.
- This part will be evaluated qualitatively from your report. Also you will use some code of this method to transform images to reduced representation for Part-2 and Part-3 of this mini-project.

## Part-2 : Naive Bayes Multi-class Classifier

- Given train images-labels and test images.
- Do PCA over train images and convert all images to a vector representation of length $N$ as per PCA. Use $N$ = 32.
- Train a Naive bayes classifier using train images and then predict the labels of test images.
- Code will run as : `python naive_bayes.py <path-to-train-file> <path-to-test-file>`
- Structure of train and test file is explained later.

## Part-3 : Linear Multi-class Classifier

- Given train images-labels and test images.
- Do PCA over train images and convert all images to a vector representation of length $N$ as per PCA. Use $N$ = 32.
- Train linear classifiers with gradient descent. You may want to use softmax to calculate loss.
- For each test image, predict its label.
- Code will run as `python linear_classifier.py <path-to-train-file> <path-to-test-file>`
- Structure of train and test file is explained below. You can generate your own train and test file from given dataset.

### Structure of train file

```
<absolute-path-to-train-1> <label-1>
<absolute-path-to-train-2> <label-2>
.
.
.
<absolute-path-to-train-N> <label-N>
```

### Structure of test file

```
<absolute-path-to-test-1>
<absolute-path-to-test-2>
.
.
.
<absolute-path-to-test-M>
```

### Your output should be as follows:

```
<test-label-1>
<test-label-2>
.
.
.
<test-label-M>
```

## Instructions

- Submit only 3 files :
  - `pca.pdf`
  - `naive_bayes.py`
  - `linear_classifier.py`
- Put these files in a folder, name it as your roll number, zip it, upload it. Thus `<roll_number>.zip` will have directory `<roll_number>` directory containing 3 files.
- You can use only `numpy` and `PIL`, no other libraries.

- Don't include any other file in your submission. Don't import from any other file.
- Don't access any other file from your code. Don't save anything by your code.
- We'll check for plagiarism. So don't copy from someone else or online.
- Not following above instructions or not following correct output format will earn you zero in assignment entirely.
- Any malpractice will lead to zero in all assignments/projects or F grade in course.