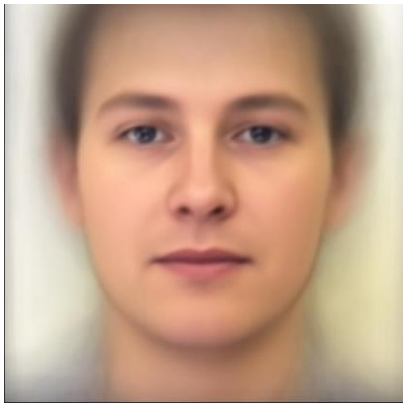
HW4

學號：r06725035 系級：資管所碩一 姓名：陳廷易

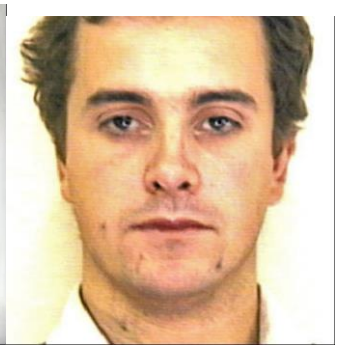
A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

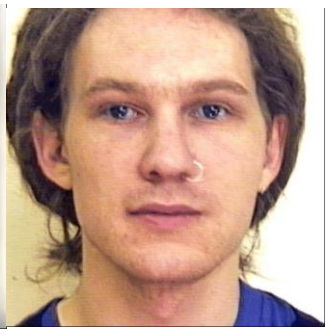


A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

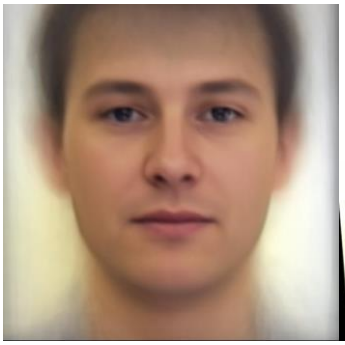
A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

▲16.jpg ▼53.jpg

▲35.jpg ▼61.jpg




A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

4.1%、2.9%、2.4%、2.2%

```
4.144624838262963 %
2.948732225112066 %
2.387711293208413 %
2.2078415569025416 %
```

B. Image clustering

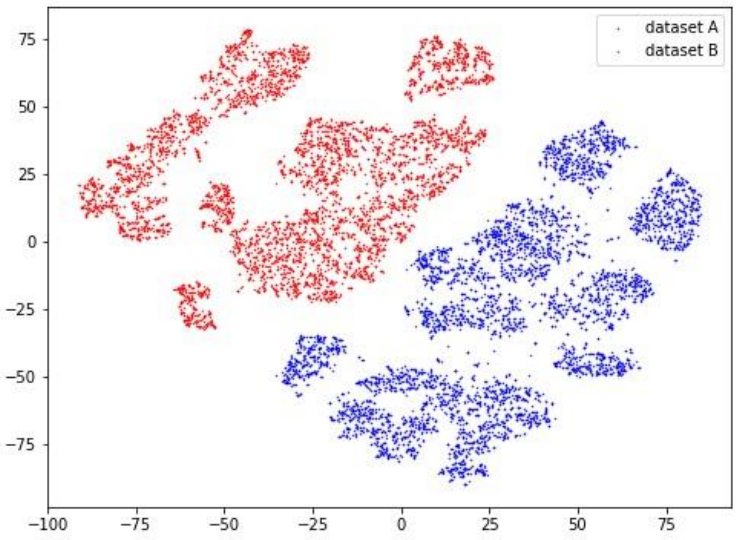
B.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

第一種 feature extraction 做法為利用 scikit-learn 的 PCA 降維，參數為 n_components=415, iterated_power='auto', whiten=True, svd_solver="full", random_state=725035。接著再使用 KMeans 方法分群參數為 init='k-means++', n_init=10, max_iter=350, precompute_distances='auto',algorithm='auto',random_state=725035, n_clusters=2，而可很完美地將兩資料集分開，於 public score 為 1，private score 亦為 1。

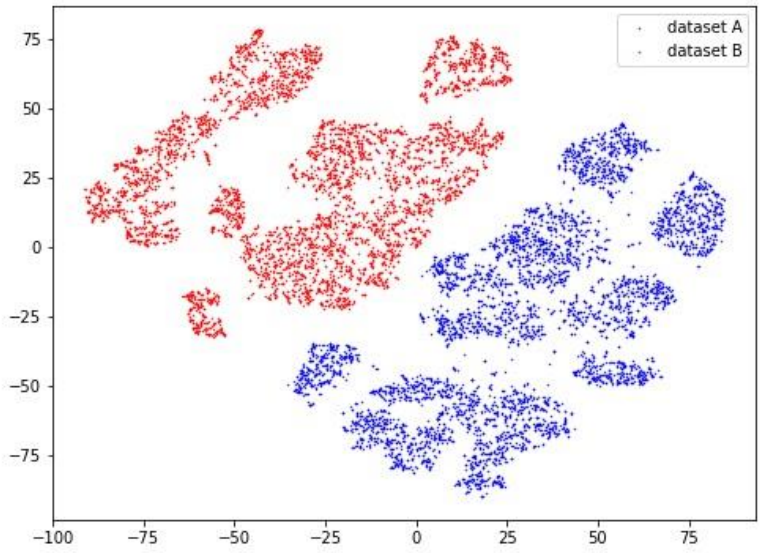
第二種 feature extraction 則利用 autoencoder 方式進行降維，訓練架構如右所示，前七層做為 encoder，取 dense_7 的輸出結果做為 code。接著再利用相同的 KMeans 參數分兩群，亦可於 public leaderboard 獲得 1 的成績，private score 亦為 1。

Layer (type)	Output Shape
input_1 (InputLayer)	(None, 784)
dense_1 (Dense)	(None, 512)
dense_2 (Dense)	(None, 512)
dense_3 (Dense)	(None, 512)
dense_4 (Dense)	(None, 256)
dense_5 (Dense)	(None, 256)
dense_6 (Dense)	(None, 256)
dense_7 (Dense)	(None, 128)
dense_8 (Dense)	(None, 256)
dense_9 (Dense)	(None, 512)
dense_10 (Dense)	(None, 784)

B.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



B.3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。



因 Precision、Recall、F1 score 皆=1 的緣故，皆經過 encoder 降維過後可看見上面兩張圖，predict 的結果與 ground truth 情況相符，且兩 dataset 中間具有相當明顯的界線。

C. Ensemble learning

C.1. (1.5%) 請在 hw1/hw2/hw3 的 task 上擇一實作 ensemble learning，請比較其與未使用 ensemble method 的模型在 public/private score 的表現並詳細說明你實作的方法。（所有跟 ensemble learning 有關的方法都可以，不需要像 hw3 的要求硬塞到同一個 model 中）

實作於 hw2 中，未使用 ensemble method 為僅使用 neural network(四層 fully-connected、relu、binary-crossentropy、rmsprop)，在 public 得分為 0.76474，而 private 得分為 0.76280。

Ensemble method 的模型為除了上述 NN model 以外，還利用 scikit-learn 裡面的 KNeighborsClassifier、DecisionTreeClassifier、SVC、LinearSVC、LogisticRegression、MLPClassifier、ARDRegression、TheilSenRegressor、SVR、Lasso、Ridge、ElasticNet、OMP 來進行訓練。訓練完成後 classifier 的結果為 0 或 1 而 regression 的結果則將>0.5 的作為 1，<0.5 的作為 0。

Ensemble method 利用上述所有模型將資料切為 75% training 25% validation，利用 validation 的準確率成績作為算術平均數的權重值。而後將所有 training data 對全部 model 重新訓練一遍，最後將預測出來的值乘上對應權重值進行 ensemble 得到最終答案。在 public leaderboard 上 score 為 0.85444 在 private leaderboard 上 score 為 0.85958，可見效果提升。