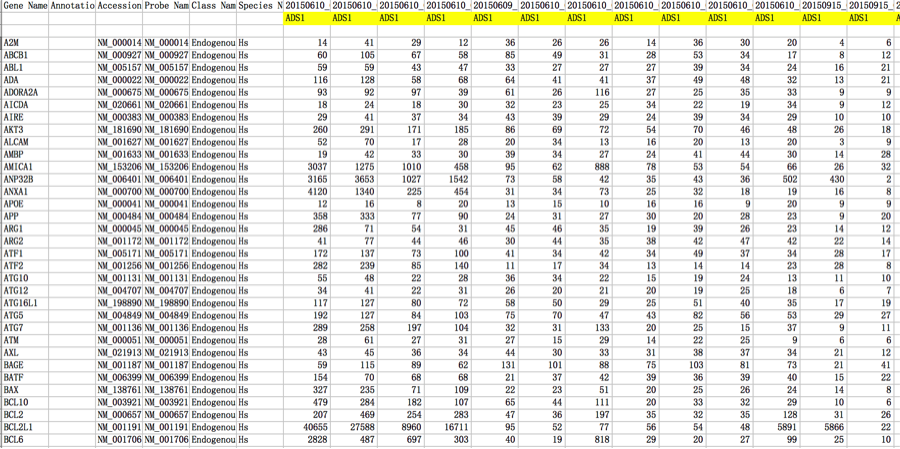
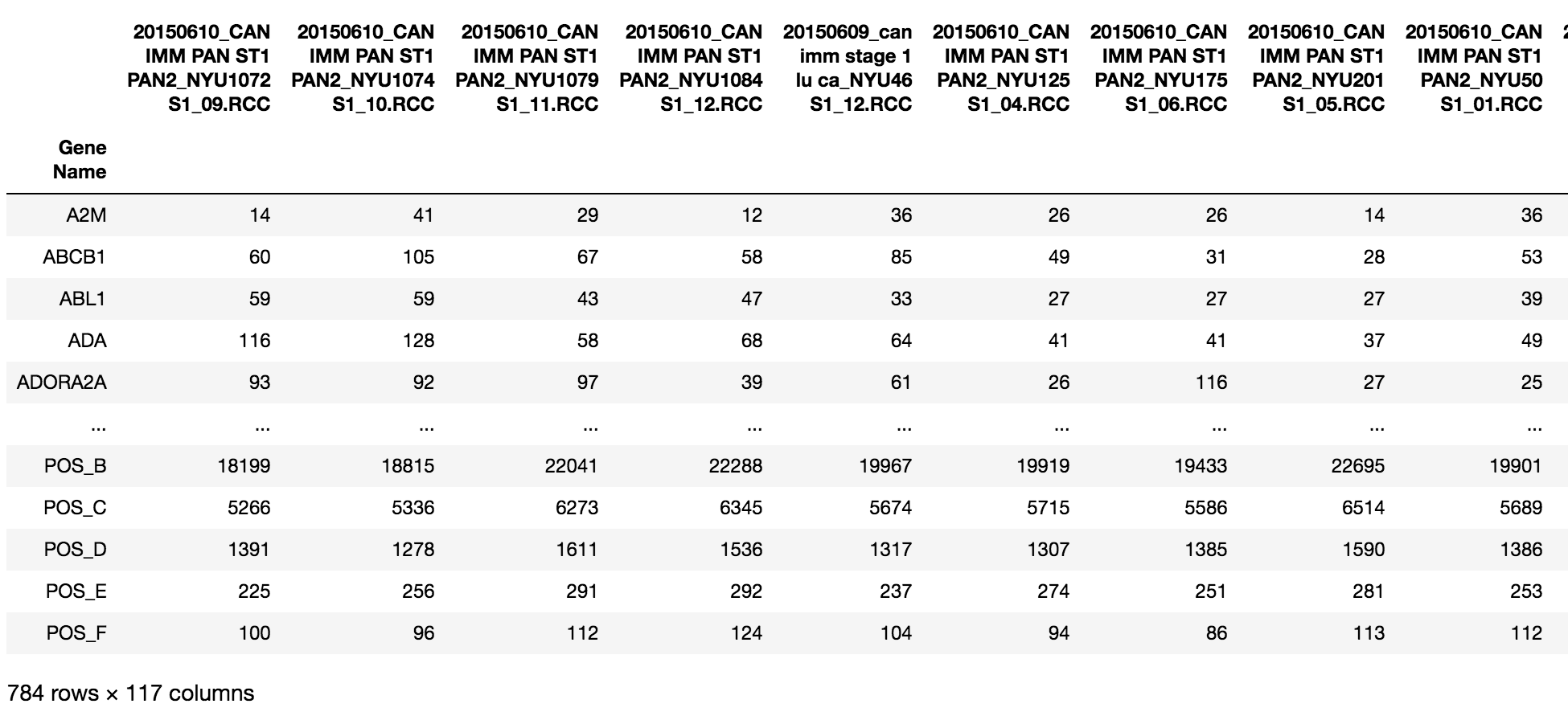
This is what the raw datas look like.



I load it in python and delete some information. Only keep the gene name, sample name (patient) and the value datas.

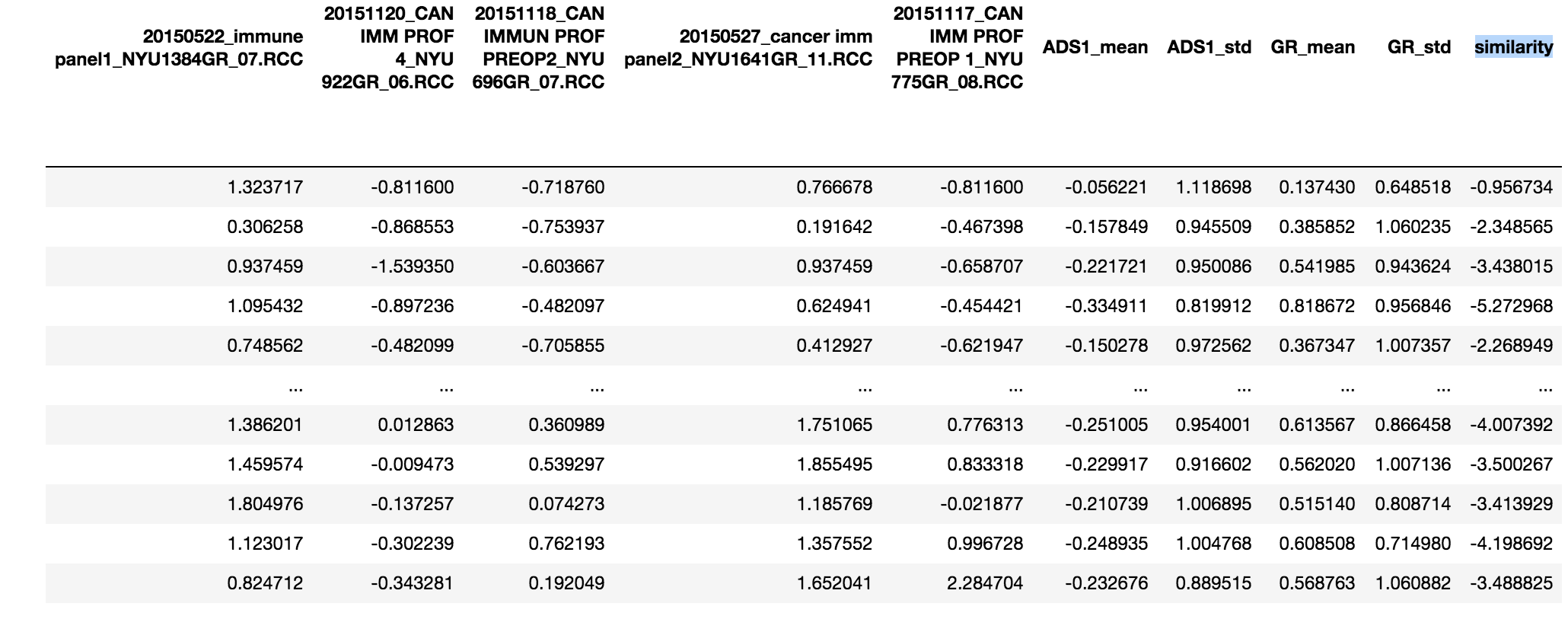


I split the dataset into train dataset and test dataset (80%train + 20%test)

And I will only use the train dataset for analysis.

93 samples in train dataset, 24 samples in test dataset.

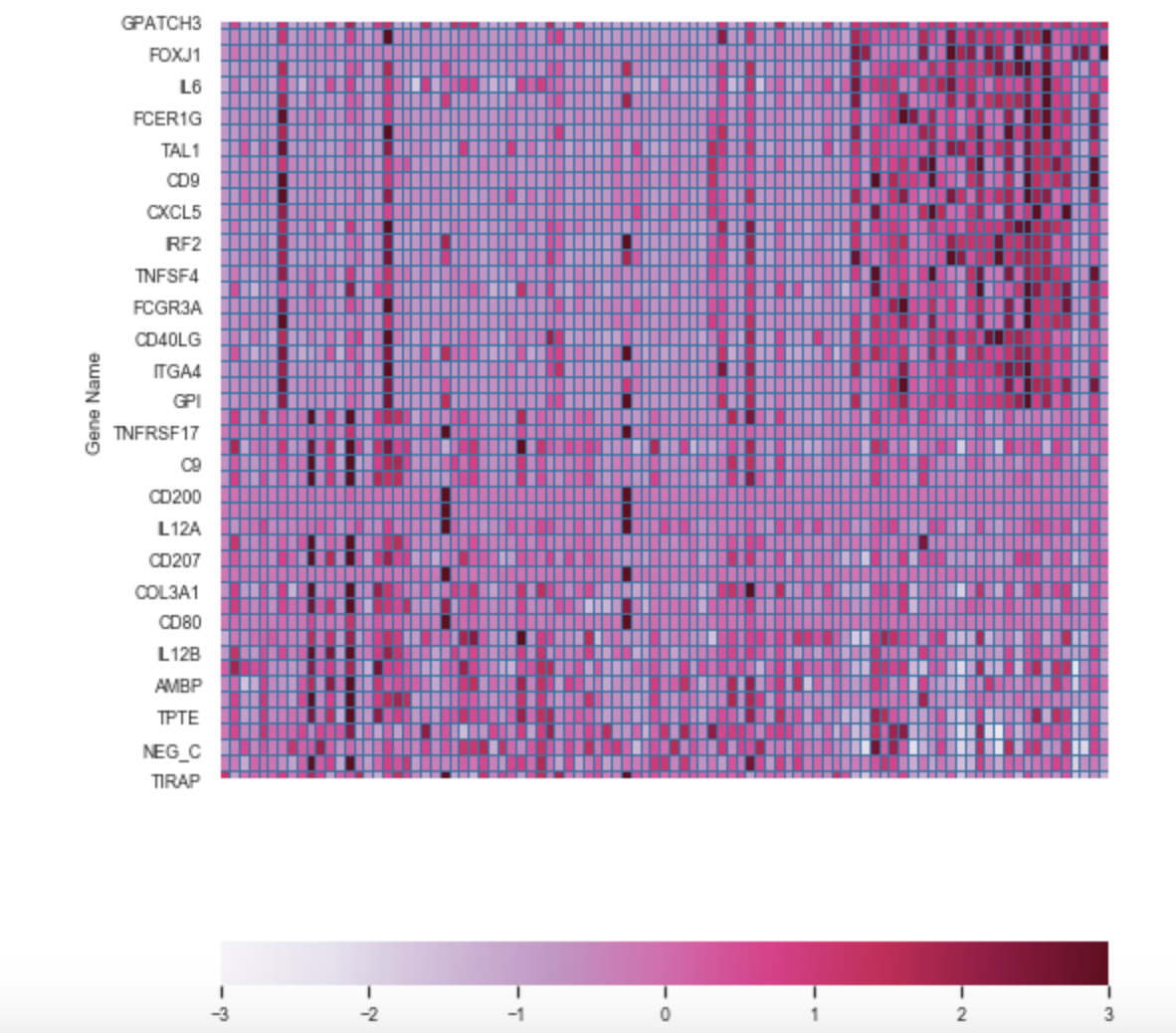
Then I do data standardisation, calculate the “similarity” of different genes based on “ADS1 and GR” group.



Then I create the heatmap based on the top 50 genes (most different expressed genes).

The left 66 columns are ADS1(abnormal), the right 27 columns are GR(normal).

The up 25 rows are most related to GR gene expression, the down 25 rows are most related to ADS1 gene expression.



Methods:

k nearest neighbors,

random forest,

naive bayes,

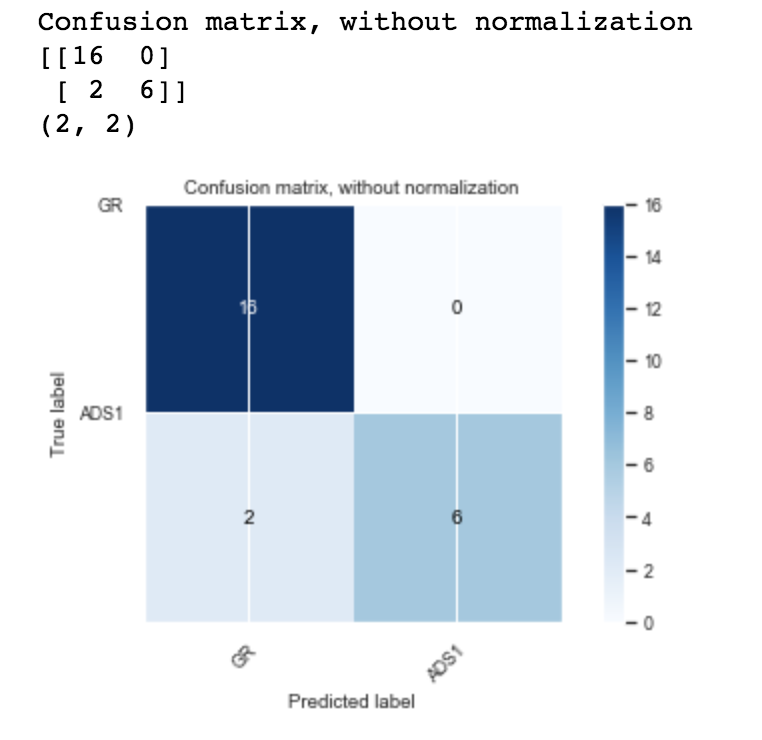
support vector machine,

logistic regression,

linear discriminant analysis

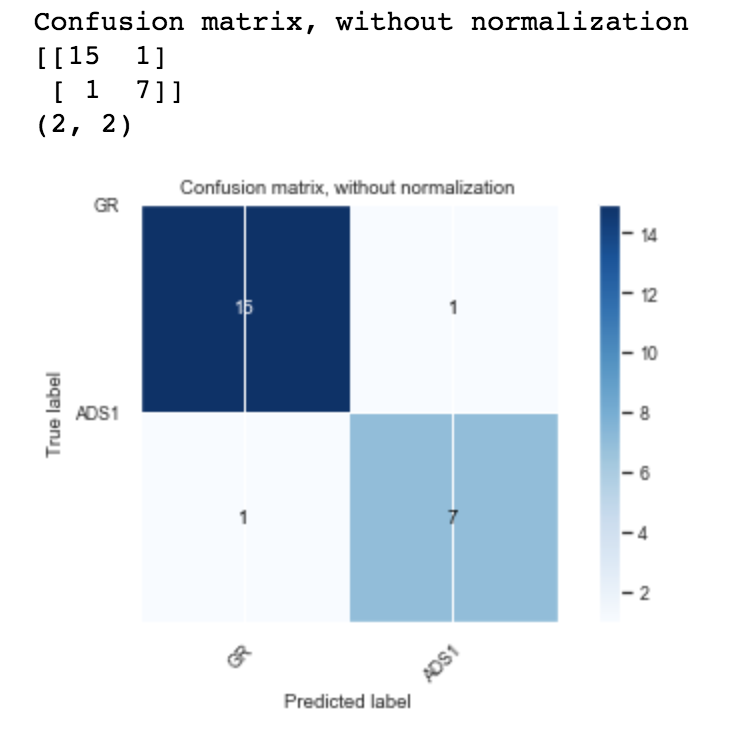
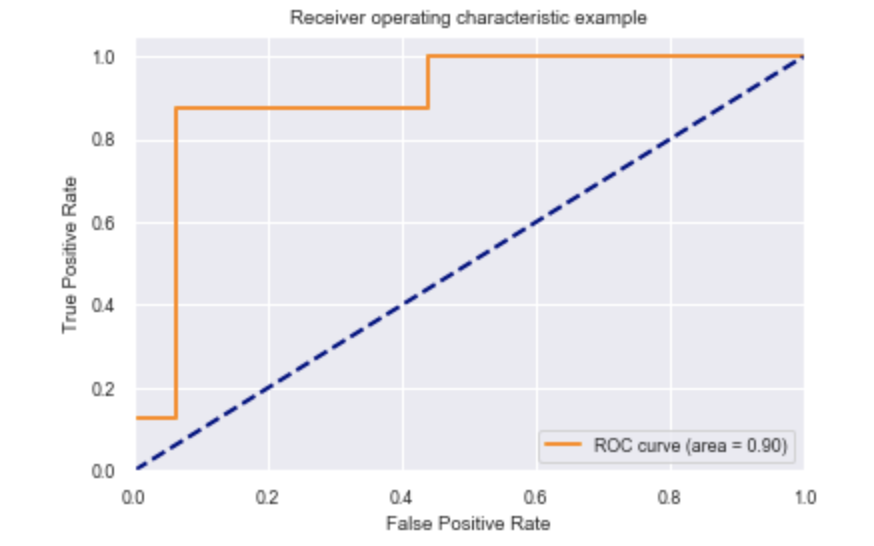
ROC curve, confusion matrix and PCA plot(red and yellow points are wrong prediction):

k nearest neighbors



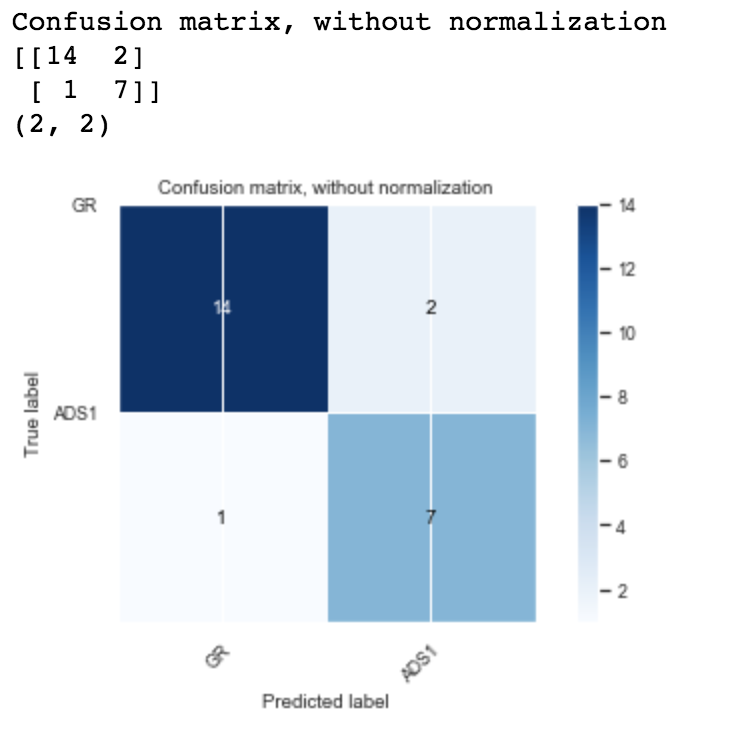
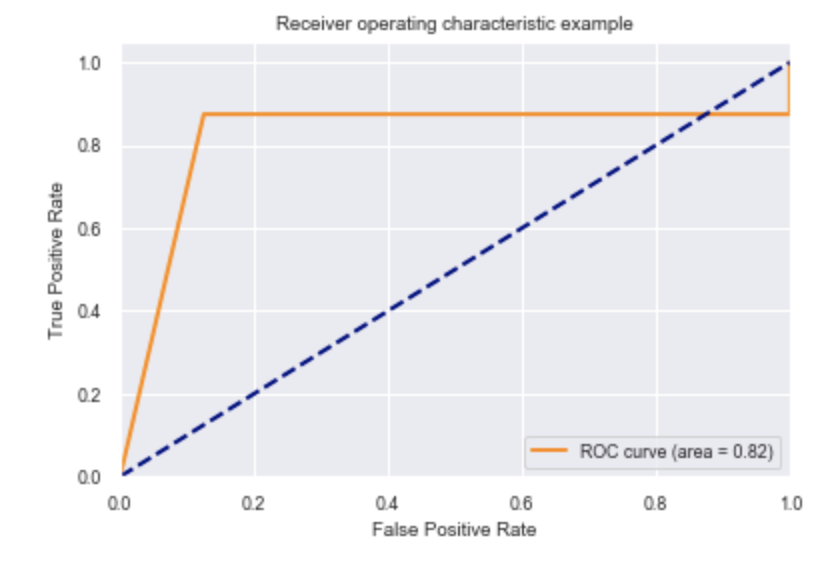


random forest



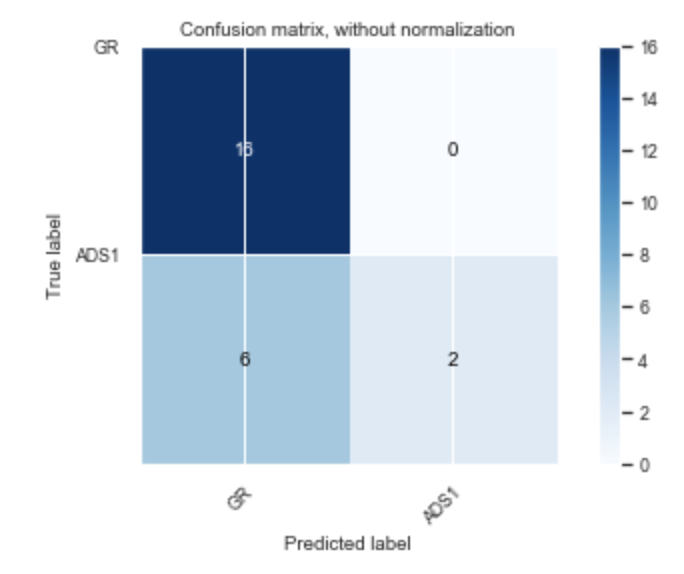
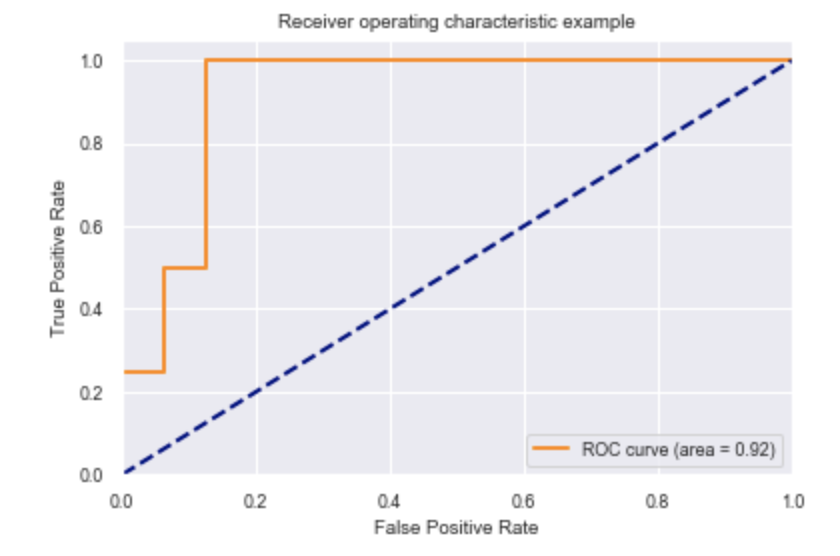


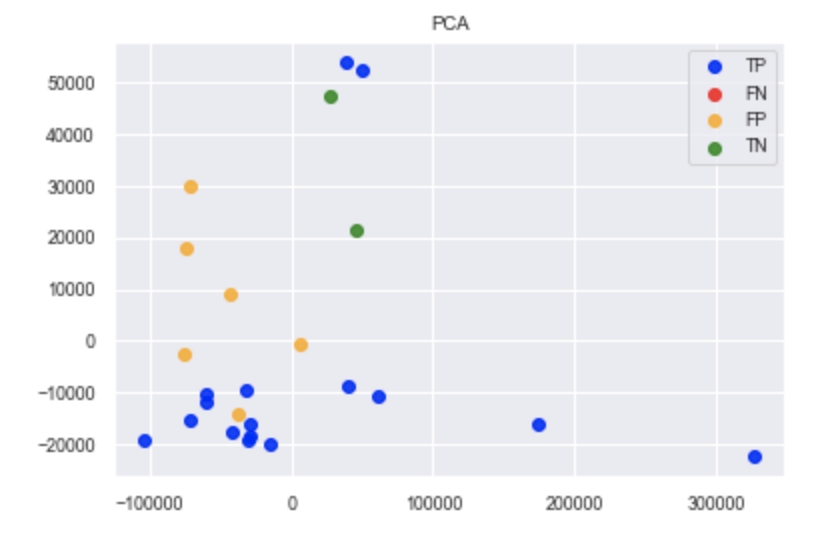
Gaussian naive bayes



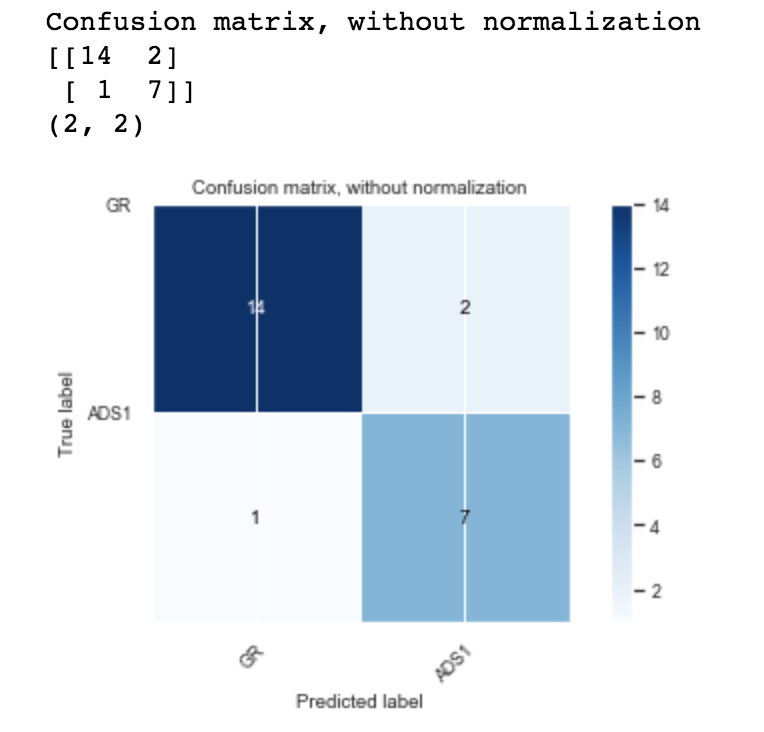
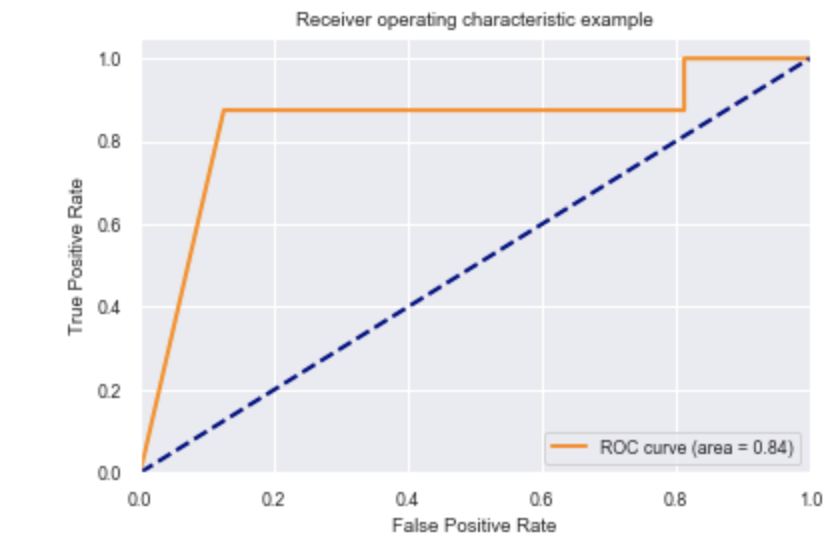


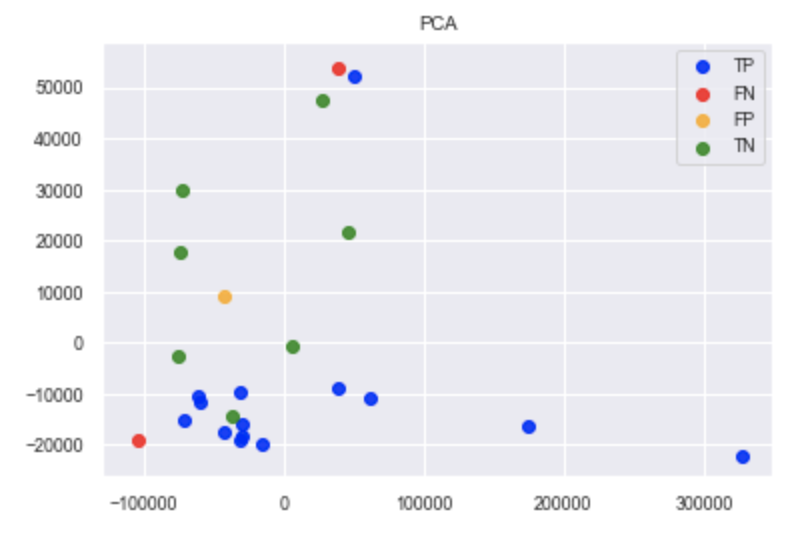
Support Vector Machine





Logistic Regression





Linear Discriminant Analysis

