Final Project

You are given the following two files.

- `sample_data_for_project_final.csv`
  This is the data table. Each column represents a sample and each row represents a variable. There are a total of 864 samples and 15,606 variables.

- `sample_metadata_for_project_final.csv`
  This file contains the labelling information for each sample. Each sample has been labelled as either 0 or 1.

Tasks

- PCA
  - Apply PCA to the data using the python scripts that you have written for one of your homework assignment this semester. Plot the scree plot and scores plot. Color the samples labelled as 0 blue and the samples labelled as 1 red.
  - Apply PCA to the data using PCA in sklearn. Plot the scree plot and scores plot. Color the samples labelled as 0 blue and the samples labelled as 1 red.
  - Compare the scree plot and scores plot you have obtained from the two methods. How similar are the two plots?

- Linear Regression
  - Apply your own linear regression scripts to the data and find all pairs of variables that are linearly correlated with correlation coefficient greater than 0.7. Make a scatter plot of all of these pairs and annotate each plot with the value of the correlation coefficient.
  - Apply linear regression to the data using linear regression in sklearn and get the R2 values.
  - Make a scatter plot of all of the R2 values greater than 0.8, if there are more than 5 of these R2 values calculated by your own scripts. The R2 values from your own scripts will be along the y-axis, and the R2 values from sklearn will be along the x-axis. Each point corresponds to a pair of variables. Compute the R2 value of the relationship between the R2 values from your own scripts and the R2 value from sklearn.

- Logistic regression
  - Split the samples into training and testing set by randomly selecting 80% of the samples for training and the remaining for testing.
  - Build a logistic regression model using your own logistic regression scripts and testing your model on the testing data set. What is your accuracy?
  - Build a logistic regression model using sklearn and testing your model on the testing data set. What is sklearn's accuracy?

- Artificial neural network (ANN)
  - Use the training and testing dataset that you have obtained when doing logistic regression.
  - Build an ANN model using your own scripts and testing your model on the testing dataset. What is your accuracy?
  - Build an ANN model using sklearn and the ANN architecture that your own ANN scripts used and testing your model on the testing dataset. What is sklearn's accuracy?

Submission

- Submit the entirety of your jupyter notebook scripts, including your own implementation of PCA, linear regression, logistic regression, and ANN and your scripts to use sklearn. Submit all of the required plots.
- Describe your results.