Project Progress Report

Title: Phenotypic Prediction from Transcriptomic Features

Team Members:

1. Jay Bhatt [111491357]
2. Faizaan Charania [111463646]
3. Jay Torasakar [111406252]

- **Objective:** To build a model that takes the Salmon output for a particular sample and predicts the original label

- **Dataset description:** We are given transcript data of 369 samples which belong to 5 different populations. More data about the transcripts and equivalence classes is available.

- **Data Preprocessing:**
    a. *Feature Selection (Extra trees classifier):* Reduced the number of features from 200,000 to 10,000 features using Extra Trees Classifier implementation of scikit-learn
    b. *Normalisation:* Z-scores are calculated for every feature of the data. Z-scores standardize the features by removing the mean and scaling to unit variance.

- **Classification models*:***
  We tried multiple approaches to classify the data and the below classifiers gave us the best results:
    a. *Random Forest*: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting
    b. *Multi Layer Classifier*: Multilayer Perceptron is a fully connected feedforward neural network which is trained using the backpropagation algorithm.

- **Model Evaluation :**
    a. The performance of our models is as follows:

Model	**F1-score (5-fold)**
Random Forest Classifier	0.8689
MultiLayer Perceptron	0.9104

    b. MLP gives us the best result, the configuration of the network is given below:
        i. Input Layer: 10000 nodes
        ii. Hidden Layer 1: 256 nodes
        iii. Hidden Layer 2: 128 nodes
        iv. Learning rate: 0.005
        v. Tolerance: 0.00001

- **Future Prospect:** In the future we will be working on the equivalence classes. Since there are too many equivalence classes right now, we are not able to accommodate all of them in the memory. So we will try to get an optimal solution so that we can work on all the given equivalence classes.